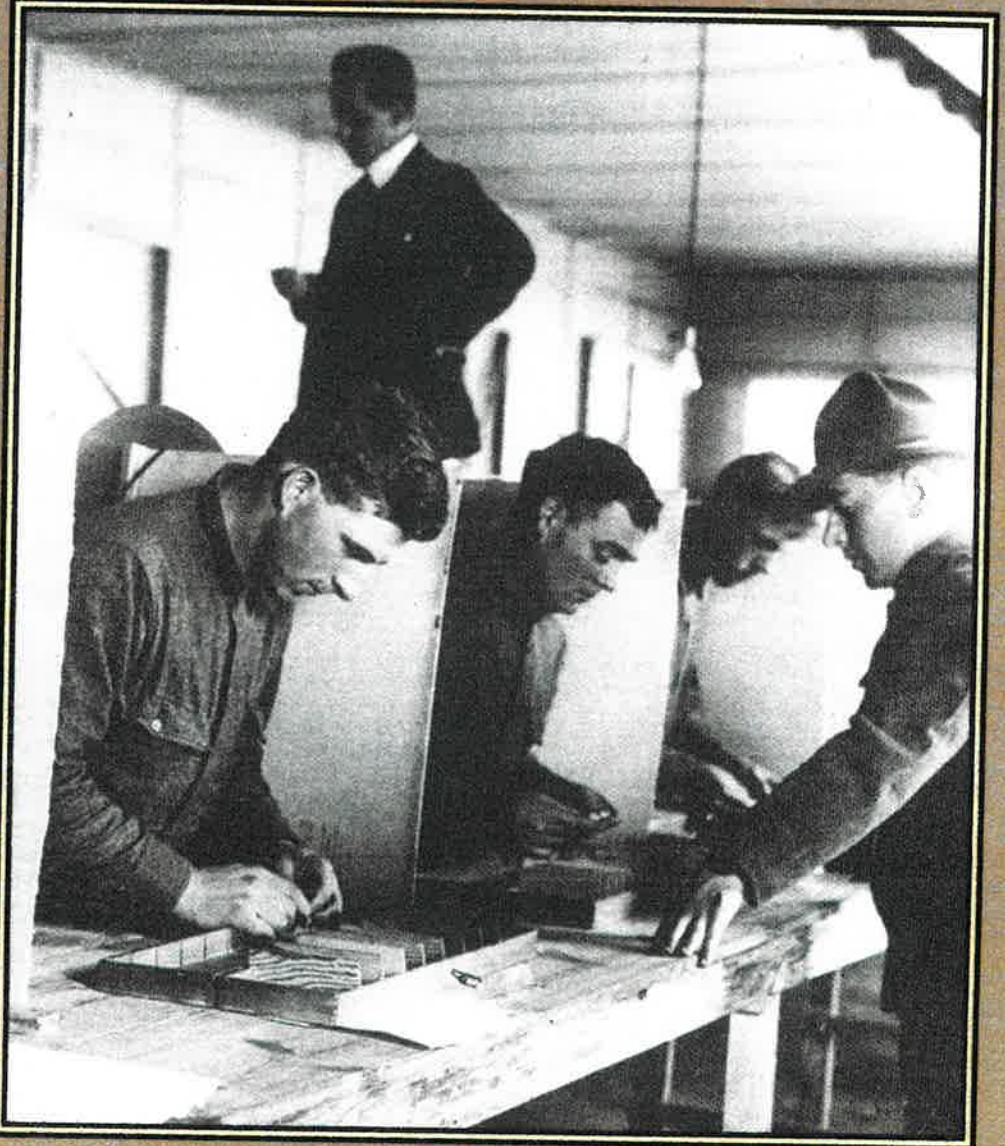
A History of the Research into Methods for Selecting and Classifying U.S. Army Personnel, 1917-2011



A History of the Research into Methods for Selecting and Classifying U.S. Army Personnel 1917-2011

A History of the Research into Methods for Selecting and Classifying U.S. Army Personnel 1917-2011

Edited by
Peter F. Ramsberger
Nikki R. Wooten
Michael G. Rumsey

With a Foreword by
Zita M. Simutis

The Edwin Mellen Press
Lewiston•Queenston•Lampeter

Library of Congress Cataloging-in-Publication Data

A history of the research into methods for selecting and classifying U.S. Army personnel 1917-2011 / edited by Peter F. Ramsberger, Nikki R. Wooten, and Michael G. Rumsey ; with a foreword by Zita M. Simutis.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-7734-2654-2 (hardcover)
ISBN-10: 0-7734-2654-X (hardcover)
1. United States. Army--Personnel management--History. 2. United States. Army--Occupational specialties--History. 3. United States. Army--Examinations--History. 4. Soldiers--United States--Classification--History. I. Ramsberger, Peter F. II. Wooten, Nikki R. III. Rumsey, Michael G.
UB337.H57 2012
355.2'230973--dc23

2012005681

hors série.

A CIP catalog record for this book is available from the British Library.

Front cover: Photo of the administration of Army Beta tests during World War I which appeared in the publication *Memoirs of the National Academy of Sciences (NAS)*, Volume XV, 1921. Source: National Academy of Sciences

The Edwin Mellen Press
Box 450
Lewiston, New York
USA 14092-0450

The Edwin Mellen Press
Box 67
Queenston, Ontario
CANADA L0S 1L0

The Edwin Mellen Press, Ltd.
Lampeter, Ceredigion, Wales
UNITED KINGDOM SA48 8LT

Printed in the United States of America

DEDICATION

This book is dedicated to generations of Army and civilian researchers who, through their commitment to scientific advancement and meeting the Army's personnel needs, have laid the foundation for a selection and classification system second to none. It is also dedicated to the millions of Army men and women who made this research possible. We extend our sincere gratitude and appreciation to soldiers who served and those who continue to serve with courage, dignity, and honor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FOREWORD

Army scientists have been in the forefront of the behavioral and social science research that underlies the selection and classification of personnel for generations. In many ways, the history of the United States Army's selection and classification research is the history of selection and classification testing in general. Innovative testing techniques developed through Army psychological research resulted in the first mass testing of individuals during World War I, which significantly expanded research and test development in the civilian personnel testing community.

That leadership has continued through the years as the Army has developed more precise ways to assess recruits' knowledge and skills based on innovations developed by its researchers. This book documents military research contributions, over a 90 year time span, leading to the important decisions that the Army makes today about whom to select and what jobs to assign to them. It also tells the story of the scientific contributions Army researchers have made over the years to the personnel testing literature and practice.

The research in selection and classification described in this volume is a case study in the development of ideas, the conduct of basic and applied research, and its continued successful application to significant and critical personnel issues. The research is placed not only in a historical context, but also in the context of current day Army selection and classification procedures.

Zita M. Simutis, Ph.D.

Director, US Army Research Institute for the Behavioral and Social Sciences, and Chief Psychologist of the US Army, retired

PREFACE

The term "research" evokes an image of test tubes, sterile laboratories, and academic isolation. Yet, the research history described in this book is something quite different. It is a story of applied research, devoted to solving specific, pressing problems that have affected hundreds of thousands of soldiers, and the security and sovereignty of our nation. This story is inextricably linked with Army history. It involves Army and civilian researchers not in isolation, but collaboratively with military leaders to identify the research problems and needs regarding the selection and classification of Army personnel. These researchers and military leaders worked together to determine the aptitudes, cognitive ability, and individual characteristics needed to be successful in Army jobs and to accomplish Army missions.

Thus, from the outset our goal was not just to describe the research methods, but also to describe the context which gave these methods meaning and purpose. This is a success story—the methods that are used today serve the Army well, just as earlier methods met the needs of earlier generations. The Army has consistently been a leader in the development and application of selection and classification methods that have been applied in other organizations. This can be attributed both to the discipline and creativity of the science involved and to the researchers' understanding of the environment in which the science was applied.

Of course, science is never a consistently uphill climb. Sometimes it moves forward rapidly, sometimes slowly, and sometimes progress is particularly hard to discern. This book describes both the hills and valleys of this progression in the history of Army selection and classification research. We

believe that through the review of this volume, not only will this part of Army history be better understood, but its utility as an exemplar of how scientific progress is achieved in general will be enhanced.

We believe that this book has something to offer those who are interested in how soldiers are selected, those who are interested in the science of selection and classification, and those who are interested in military history generally. We wish to express our thanks to the chapter authors and the many others who made this book possible.

ACKNOWLEDGMENTS

Four individuals not listed elsewhere deserve special thanks for their contributions to this effort. Deirdre Knapp provided major assistance to the planning and execution of this project. Dorothy Young was instrumental in identifying and making available many of the references cited. LaVonda Murray provided much-needed formatting assistance in the editorial process. Eric Gade created the Topic Index.

Support for this effort was provided by the U. S. Army Research Institute for the Behavioral and Social Sciences, Department of the Army, Delivery Order 33, Contract Number DASW01-03-D-0015. However, the opinions, interpretations, conclusions, and recommendations expressed in this volume are those of the authors, and do not necessarily represent those of the Department of the Army or the Department of Defense.

PART I

INTRODUCTION AND OVERVIEW

CHAPTER 1

PUTTING IT IN PERSPECTIVE

Peter F. Ramsberger
Human Resources Research Organization (HumRRO)

> ***Wars may be fought with weapons, but they are won by men.***
>
> General George Patton

From the beginning of time to the present day, the major determinants of the outcome of any military operation are the people who carry it out. No matter how sophisticated the equipment or techniques that are brought to bear, if they cannot be effectively used or implemented by the men and women so charged, the ultimate result will be failure. Of course, leadership also plays a key role, whether it is higher-level commands making key strategic decisions, or leaders on the ground guiding troops in the heat of the battle.

Throughout much of its history, the technical demands the United States Army put on those who served were limited in scope. Although soldiers were often required to endure harsh conditions as they fought the enemy, to be successful the men had to know how to fire a rifle, operate a cannon, move and fight in formation, or perhaps ride a horse—but typically not much more. As technology advanced, so too did the range of tasks that soldiers could be asked to learn. Radio operation and repair became essential to battle efforts. With the introduction of aircraft into the Army's combat arsenal, aviators had to

master a wide range of skills to successfully carry out their missions. In today's Army there are few jobs that have not been touched by technological advances in some way. A recent study of the conditions under which soldiers of the 21st century will be expected to operate anticipated the continuation of this trend (Sager, Russell, Campbell, & Ford, 2005). It describes a learning environment in which there will be a "greater requirement for continuous learning and the need to independently maintain/increase proficiency on assigned tasks" (p. 26).

Throughout history, the size of the Army ebbed and flowed depending on the specific needs at any given point in time. Starting with the colonists, there has been a general mistrust of large, standing armies, based on the concern that the military could usurp civilian authority. Thus, in times of war every effort was made to muster men to take up arms in defense of the nation or cause, but once hostilities ceased the Army was allowed, and even encouraged, to return to pre-war numbers. With small armies and limited roles within them, the process of selecting soldiers and assigning them to jobs was largely done on a one-on-one basis. Who better to decide the job given a soldier would have than the leaders who got to know their men personally basis and thus determine in what capacity he should serve.

Although the size of the Army did decrease substantially following the conclusion of World Wars I and II, it never again was an institution that could be managed on the basis of personal acquaintance. According to figures provided by Weighley (1984), the number of soldiers, both officer and enlisted, reached a nadir in 1950, when just under 600,000 men and women were in uniform. In the early 1950s with the threat of the Soviet Union and a return

to the draft, this number swelled almost three-fold. The Army's strength remained high throughout the 1950s and 1960s, and only began to decline as operations in Vietnam gradually ended. From that time until the drawdown of the early and mid-1990s, force levels remained relatively stable at between 600,000 and 700,000. The effect of the drawdown was to reduce this number to under one-half million, supplemented by a nearly equal number of Army National Guard and Reserve forces. Even with these reductions, however, the need for a systematic and effective method of selecting those who will serve, and assigning them jobs where they will be able to perform effectively, remained great.

Another piece to the puzzle had to do with the advancement of personnel psychology. As America moved from a largely agrarian nation to one with an expanding manufacturing base, the diversity of skills required to support the economy grew exponentially. Large industries began to discover the need to have systems that would allow them to hire the most able people for the particular tasks to be performed, and assign them based on other than an ad hoc basis. In the years leading up to World War I, the field of personnel psychology and the acceptance of its methods became more widespread. The use of tests to identify those who could serve as clerks and typists, managers and supervisors, began to become more common practice, and the techniques within the field more advanced. With the onset of the war, where better to try to apply these methods than in what became the largest organization in the nation, the United States Army?

As described in Chapter 7 of this volume, another factor that influenced the selection and classification system was the move to the All Volunteer Force (AVF) in 1973. One of the prime inducements to encourage young people to volunteer for Army service is the promise of military and job training that will

benefit them in their post-service lives. Such an enticement is greatly enhanced if potential recruits can have input in selecting the training they will receive. This provided further support for the decision to move to a single test used by all Services for both selecting qualified youth and determining the occupations for which they are best suited. As a result, an interested young man or woman could be given occupational options at the time of enlistment, along with assurances that their preference would be honored down the line.

In the past, the Army had tried to capitalize on the skills that men brought with them to service. Particularly during periods of conscription, many new soldiers would leave jobs back home that had direct military counterparts. As described in Chapter 4 of this volume, such efforts were formalized in World War I, with a systematic process put into place to learn about the talents of those who stepped forward (or were required) to serve so that they could be of the greatest effectiveness in the war effort.

In today's Army, however, the situation is vastly different. In 2006, 70 percent of new active duty recruits were 21 years of age or younger on entry. The typical recruit is a high school diploma graduate, with little or no work experience that can be gainfully put to use in the Army. Given that the most popular enlistment option is for a three-year term, it is clear that new recruits need to have the capacity to absorb training quickly so they can join their units and begin functioning as part of the team. With the increasing complexity of military occupations, the need to find youth who can excel in this regard, and to place them in a job to maximize their chance for success, is of critical importance.

Another impact of the AVF was that it substantially increased the investment made by the Department of Defense (DoD) and the Services in obtaining manpower. Advertising, recruiter time, enlistment bonuses, and other incentive programs had to be put into place in order to inform American youth about the opportunities available to them in the military, and to persuade them that this option will be of benefit to them, as well as their country. As discussed in Chapter 8, the costs are high when recruiting efforts result in the accession of those who do not become productive soldiers. Whether losses are due to attrition for behavioral reasons or training failures, the ramifications in terms of resources, unit morale, and performance can be great.

For this and other reasons, significant effort has been put into developing methods of identifying who among the vast number of applicants for military service are most likely to succeed. Such research repeatedly demonstrated strong linkages between performance on the military selection and classification test, the Armed Services Vocational Aptitude Battery (ASVAB), and subsequent training outcomes. Chapter 8 describes a large-scale, joint-service project aimed at taking this a step further by establishing that ASVAB scores are also related to performance on the job, thereby eliminating doubts about the value of the ASVAB in selecting individuals who can be effective soldiers. Research is also described that shows success, particularly as indicated by completion of a term of service, is positively related to education level and credential type. Those who fail to receive their high school diploma are much more likely to also fail to complete their enlistment obligation than individuals with alternative credentials or no credentials at all.

In summary, a number of developments over time have led to the creation and evolution of selection and classification processes

within the Army to ensure a highly qualified force, including: (a) the establishment of permanent, large manpower base; (b) the increasing cognitive demands placed on soldiers at all levels; (c) the move to an all-volunteer force composed of young, largely inexperienced new recruits; (d) the increased expense associated with recruiting and training those who volunteer to serve; and (e) advances in the field of employment selection and classification that allow for the application of sophisticated scientific methods to improve screening and assignment, and thereby reducing costs associated with training failures and other forms of attrition.

The Army and other military services offer a unique venue in which to study personnel selection and classification. Both functions are carried out on a regular basis and in large scale. Because of the critical nature of the task, it has high visibility and, given the costs associated with unsuccessful outcomes, funding for efforts to develop and improve screening and placement methods has generally been made available. As a result, the case can easily be made that the military has been at the forefront of selection and classification research and development and has contributed as much as any institution to the advancements in these processes. This volume is intended to provide a history of that work, from its earliest days of simple, relatively crude, paper-and-pencil tests, through the present day use of methods such as computer-adaptive testing.

In Part I, we follow this introductory chapter with a focus on selection and classification practices as they are carried out today, including criteria that are not discussed elsewhere in the book, such as educational and moral standards.

In Part II, we start with a brief overview of the means and methods used to populate the United States Army from the days of the first settlers up to the first World War. This provides a

context for what follows and highlights patterns in manpower practices that, although acceptable throughout much of the country's history, are untenable in times of relatively constant threat and the attending need for continuous vigilance. The next three chapters, Chapters 4, 5 and 6, address work done in World War I, World War II, and the post-war period, respectively. These were periods of major advancement and, particularly in the early stages, an important time for personnel psychologists to convince military leaders that psychologists did, indeed have an important contribution to make to achieving the mission. Chapters 7 and 8 focus on equally important eras, as the nation moved to an All Volunteer Force and new questions arose, and were answered, regarding the value of selection and classification research and development. This included what remains today the largest effort ever undertaken to demonstrate the validity of selection tests in predicting success in the Army—Project A. This is followed by a detailed historical analysis of the Army's efforts to implement competency testing for soldiers, both as a means to identify the abilities of the force as well as those who will rise within its ranks. Chapter 9 focuses specifically on classification research and the advances that have been made over time in developing measures that allow us to determine who is most suitable for specific Army jobs, even with little information from an individual's past experience.

Part III of the book is devoted to selection research and development focused on specific Army occupations and needs. These include the selection of officers, the identification of soldiers likely to succeed in Special Forces units, and the selection of aviators. In Part IV, we try to place the volume's contents in perspective, by looking ahead at what we can expect in the field, and also looking back to provide the context that allows us to understand the tremendous contributions made by

thousands of researchers, military manpower specialists, and soldiers to the field of personnel selection and classification.

References

Sager, C. E., Russell, T. L., Campbell, R. C., & Ford, L. A. (2005). *Future soldiers: Analysis of entry-level performance requirements and their predictors* (Tech. Rep. No. 1169). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Weighley, R. F. (1984). *History of the United States Army*. Bloomington, IN: Indiana University Press.

CHAPTER 2

ARMY SELECTION AND CLASSIFICATION TODAY

Rodney A. McCloy
Human Resources Research Organization (HumRRO)

Each year, the U.S. Department of Defense (DoD) enlists tens of thousands of recruits into the Military Services. In FY 2006, approximately 167,000 recruits donned a uniform for the first time and joined active (as opposed to Reserve or National Guard) components (U.S. Department of Defense, 2007). Nearly 70,000 of these new recruits accessed into the U.S. Army.

The Services strive to recruit, enlist, and retain the best talent available. To identify such recruits, the Services collect various types of information on applicants using measures of such characteristics as cognitive ability, education, physical abilities, medical history, and conduct. An applicant's standing on these characteristics helps determine his/her qualification for military service. Having selected a cohort of applicants for enlistment, each Service then performs various activities regarding classification—the assignment of recruits to military occupations. Information used to select applicants is standard across all of DoD, but classification/assignment procedures differ by Service.

This chapter describes the current selection and classification procedures of the U.S. Army. After providing an overview the types of measures used for selection and classification, the

chapter concludes with a brief discussion of potential forthcoming modifications to the Army's selection and classification process.

Measures of Individual Characteristics

DoD considers an expansive array of information when evaluating an applicant's qualification for military duty. This section describes the types of data collected and the measures used.

Age

Federal law (10 U.S.C., § 505) establishes the minimum (17 years—parental consent required) and maximum (42 years[1]) age for military enlistment. This maximum age is not uniform across Services/components. Rather, DoD permits each Service to establish its own maximum age limits for non-prior service enlistments. Service-specific limits are typically lower than the DoD maximum (e.g., the maximum age for the Active Air Force is 27 years), although the Army has adopted the DoD maximum of 42 years. Applicants with prior service history can receive a waiver to enter at an age above the maximum. The extent to which the applicant's age can exceed the maximum age varies by Service.

The majority of recruits in any given enlisted cohort range from 18 to 23 years of age. For example, 84.3% of the FY 2006 non-prior service active component enlisted accessions were 18-23 years of age,[2] and nearly two-thirds (64.3%) were 18-20 years of age (U.S. Department of Defense, 2007). Research has shown that older recruits are more likely to complete their

[1] The maximum age had been 35 years until Congress raised it in 2006.
[2] The norming population for the AFQT comprises a national population of youth in this age range.

enlistment terms than are younger recruits (Hooper, Paullin, Putka, & Strickland, 2008).

Cognitive Ability: The Armed Services Vocational Aptitude Battery (ASVAB)

To determine general trainability and future performance potential, DoD administers the Armed Services Vocational Aptitude Battery (ASVAB) to all applicants. The ASVAB assesses standing on general cognitive ability and various technical aptitudes.[3] The ASVAB comprises nine tests: (a) Word Knowledge (WK), (b) Paragraph Comprehension (PC), (c) Mathematics Knowledge (MK), (d) Arithmetic Reasoning (AR), (e) General Science (GS), (f) Auto and Shop Information (AI), (g) Mechanical Comprehension (MC), (h) Electronics Information (EI), and (i) Assembling Objects (AO). The GS, AS, MC, and EI tests are considered technical tests and are used primarily for classification/placement decisions. The AO test, which is a measure of cognitive ability (specifically, nonverbal reasoning), remains somewhat experimental in nature, being administered regularly but appearing in just one Aptitude Area Composite—the Navy's Mechanical Maintenance 2 (MEC2) Composite.

DoD combines the verbal (PC, WK) and mathematical (AR, MK) ASVAB tests into a composite measure of general cognitive ability (often termed "trainability")—the Armed Forces Qualification Test (AFQT). The AFQT score is computed as a sum of standardized test scores:

$$AFQT = AR + MK + 2(WK+PC) = AR + MK + 2VE,$$

where VE is the Verbal composite score resulting from computing the sum of standard scores on the WK and PC tests.

[3] A detailed history of the ASVAB is provided in Chapter 3 of this volume.

The AFQT, which is generally reported as a percentile rank score, serves as DoD's enlistment test. AFQT scores are also frequently reported according to the percentile category in which they fall (see Table 3.1). Percentile scores are based on a distribution of AFQT scores obtained by a norm group comprising a representative sample of youth aged 18-23 (with oversampling of African American and Hispanic youth; Segall, 2004).[4] By law, non-graduates in Category IV and all in Category V are ineligible to enlist.

Table 3.1. AFQT Categories and Their Corresponding Percentile Score Ranges

Category	Percentile Score Range
I	93 – 99
II	65 – 92
III-A	50 – 64
III-B	31 – 49
IV-A	21 – 30
IV-B	16 – 20
IV-C	10 – 15
V	1 – 9

Research has demonstrated time and again that recruits with higher AFQT scores perform better in job training (U.S. Department of Defense, 1984) and on the job (Oppler, McCloy, & Campbell, 2001; Oppler, McCloy, Peterson, Russell, &

[4] The norms based on this 1997 youth population went into effect on 1 July 2004.

Campbell, 2001). The performance advantage of recruits in higher AFQT categories continues through time. For example, it takes nearly 3 years for recruits scoring in Category III-B to reach the average performance levels achieved within the first few months on the job by recruits scoring in Categories I and II (U.S. Department of Defense, 1992). Relations to other important outcomes, such as attrition, have also been demonstrated (Hooper et al., 2008; McCloy & Putka, 2005).

Recruit Quality

Applicants scoring at or above average (the 50th percentile) on the AFQT are in Categories I – III-A (see Table 3.1). DoD considers such applicants who also possess a high school diploma to be *high-quality applicants*. DoD has established recruit quality requirements that each Service must meet. These requirements help DoD forge enlistment policies and establish recruiting budgets. Of particular import to the latter is the tradeoff between recruiting cost and recruits' job performance. High-quality recruits cost more to attain than do lower-quality recruits, but they also are more likely to complete their enlistment terms and perform better in training and on the job.

For the past 15 years, DoD has used a cost-performance tradeoff model (CPTM; McCloy et al., 1992; P.F. Hogan & Harris, 1994; Smith & P.F. Hogan, 1994) to help establish and monitor recruit quality standards. Recruit quality standards are defined in terms of the percentage of the force having high school diplomas (90%) and scoring in AFQT Categories I-IIIA (60%).[5]

[5] *Editor's note.* It should be noted that this use of the term "quality" is rooted in a particular historical and policy context. Recently, the Army has preferred to view high school diploma status and AFQT scores as measures of potential rather than quality. This topic is discussed further in Chapter 14.

Foreign Language Recruiting Initiative (FLRI)

Changing demographic trends in the United States have led the Army to design and evaluate the Foreign Language Recruiting Initiative (FLRI). FLRI rests on the notion that the AFQT does not measure the true aptitude of non-native English speakers. To address this difficulty with identifying the cognitive ability of such applicants, the FLRI program identifies qualified non-native English-speaking individuals, enlists them, provides them with English-as-a-Second-Language (ESL) training, and then tests those who graduate ESL on the Armed Forces Classification Test (AFCT) – the in-service version of the ASVAB. This post-ESL AFQT score informs Army decisions regarding recruit retention, as well as job assignment. FLRI recruits whom the Army retains after ESL training go on to Basic Combat Training (BCT).

The Army conducted three evaluations of the FLRI program (DiFazio, McCloy, Hooper, & Russell, 2008; DiFazio, Putka, Wasko, McCloy, & Shaw, 2010; Medsker, Le, & Knapp, 2005). The final evaluation (DiFazio et al., 2010) concluded that (a) the test used to screen FLRI applicants (the Assembling Objects test from the ASVAB) was a valid predictor of several valued FLRI outcomes (e.g., post-ESL AFQT scores, scoring in a higher AFQT category after ESL training) and (b) both recruiters and participants view the program positively. In addition, a project memorandum for record (Ani DiFazio, personal communication, October 22, 2010) reported that 77% of participants who entered the program under its current configuration (a) improved their post-ESL AFQT category to IIIB or higher, and (b) exhibited lower attrition rates (12-month rate = 13.2%; 36-month rate = 19.5%) than did Regular Army

accessions in FY04 through FY08 [12-month rates = 13.2% - 20.1%; 36-month rates (FY04-FY07 only) = 28.6% - 32.5%].

Education

The military services do wish to recruit high-performing personnel, but it is also important that recruits complete their enlistment terms. Recruits with high cognitive ability (who would be expected to perform at a high level) provide little return on investment if they leave the military well before their enlistment term is to end. The Services give differential enlistment priority to applicants, depending on their level of education. The Services place particular value on those who have a high school diploma, because (as cited below) research has shown that high school diploma graduates are far more likely to complete their first enlistment term (i.e., are less likely to attrite) than are recruits who lack a high school diploma.

Attrition has significant monetary implications. A General Accounting Office report to Congress (General Accounting Office, 1998) stated that each attrition incident costs DoD in excess of $35,000. Assuming a first-term attrition rate of 30% across the Services' non-prior-service accessions (a conservative value) numbering approximately 167,000 (U.S. Department of Defense, 2007) leads to attrition costs in excess of 1.75 billion dollars—estimates that others have argued are conservative (Brose, 1999; McCloskey, 1999). A recent memorandum from the U.S. Army Training and Doctrine Command (TRADOC) stated that the cost of training a Soldier from the beginning of the recruiting process until that Soldier reaches his/her duty station averages about $50,000 (J. Thomas, personal communication, February 8, 2005). Use of this estimated cost per attriting Soldier would lead to annual attrition costs in excess of 2.5 billion dollars.

Therefore, identifying recruits who are likely to complete their enlistment term and thereby provide the Services with returns on their training and personnel costs is not a trivial matter. Fortunately, myriad studies have demonstrated that recruits who possess a high school diploma are much more likely to complete their first enlistment term than are recruits who lack a diploma or possess an alternate education credential, such as a GED (Buddin, 2005; Ellis, 1999; Elster & Flyer, 1981; Flyer, 1956; Knox, 1998; Laurence, 1997; Laurence, Ramsberger, & Arabian, 1996; Lindsley, 1995; Ramsberger, Laurence, McCloy, & DiFazio, 1999). In addition, a recent, comprehensive analysis of first-term Army attrition again identified educational status to be one of the best predictors of first-term attrition (Strickland, 2005).

For reporting purposes and for determining personnel policy, DoD breaks education level into three classes, or "tiers." Tiers are ranked in terms of their desirability, from most to least desirable:

- *Tier I* primarily includes applicants who are high school diploma graduates, but also comprises those successfully completing at least one semester of full-time enrollment in college (i.e., those earning credits in an institution recognized by the Accredited Institutions of Post-secondary Education) as well as those obtaining an Adult Education Diploma, which DoD defines as "A secondary school diploma awarded on the basis of attending an alternative, continuation, adult, or charter school/program" (United States, 2010, paragraph 2-2c, p. 6), the distinguishing feature being that attendance in the program compares favorably with attendance in high schools.

- *Tier II* comprises those holding alternative credentials (e.g., Graduate Equivalency Diploma, Correspondence School Diploma, Certificate of Attendance). Attrition rates in these programs are higher than those among high school graduates.
- *Tier III* comprises those applicants who lack a recognized credential.

The Services limit the number of applicants who can enlist from Tiers 2 and 3. For example, the Army allows up to 10% of its recruits to come from Tier 2 (consistent with DoD's recruit quality benchmark of 90% high school diploma graduates—see discussion at the end of the next section), but all such applicants must score in or above AFQT Category IIIA. Although Tier 2 applicants have higher attrition rates than do those in Tier 1, about half of Tier 2 applicants complete their enlistment term. The Army has been using the Assessment of Individual Motivation (AIM; White & Young, 1998)[6] to help identify those Tier 2 applicants who might be at lowest risk for attrition (White et al., 2004).

Recent Army Programs Involving Education Credentials

Until 1998, DoD considered homeschoolers as Tier 2 recruits, thereby equating them with individuals who hold an alternate education credential. This categorization changed for a 5-year period (1998-2003), during which homeschooled applicants entering the military were treated as Tier 1 applicants. An evaluation of that program found their attrition rates to more closely mirror the rates of Tier 2 recruits (Wenger & Hodari, 2004), and thus homeschoolers returned to the Tier 2 designation. Attrition rates for homeschoolers were low through

[6] The AIM is discussed further in the final section of this chapter.

the first enlistment year but rose during the second and third years. Attrition rates for homeschooled recruits who scored in AFQT Categories I - IIIA, however, closely resembled rates for high school diploma graduates. On 1 June 2007, DoD initiated a program to categorize homeschoolers scoring in AFQT Categories I – IIIA as Tier 1 recruits. This trial program will last through 30 September 2011 (United States, 2010, paragraphs 2-3d, pp. 8-9).

Another recent Army development involves a new 4-week course designed to help applicants who qualify on other enlistment requirements to receive a GED (Association of the U.S. Army, 2008). The course is offered by the new Army Preparatory School at Fort Jackson, SC. Should the program prove successful, the Army might extend the course length to 8-10 weeks.

Medical Status

Applicants must complete and pass a series of medical screens at the Military Entrance Processing Stations (MEPS). Medical screeners use the Military Physical Profiles Serial System to evaluate the medical fitness of each recruit with regard to six categories, each affiliated with a body "system" and together constituting the acronym "PULHES" (Sackett & Mavor, 2006):

- P – Physical capacity or stamina (general physical capacity, including cardiopulmonary, nervous, and gastrointestinal systems, as well as allergies and endocrine/metabolic disorders)
- U – Upper extremities (strength, range of motion, and general efficiency of hands, arms, shoulder girdle, upper back)

- L – Lower extremities (strength, range of movement, efficiency of feet, legs, lower back, and pelvic girdle)
- H – Hearing and ears (auditory acuity, diseases/defects of the ear)
- E – Eyes (visual acuity, diseases/defects of the eye)
- S – Psychiatric (personality, emotional stability, and psychiatric diseases).

Screeners assign a score of 1 to 4 for each category, with 1 indicating a high level of medical fitness, 2 indicating a medical/physical condition that may impose some limits on activities, 3 indicating one or more medical/physical conditions that may result in significant limitations of military duty, and 4 indicating one or more medical/physical conditions so severe as to greatly constrain performance of military duties.[7] Few jobs can accommodate individuals with scores of 3 or 4 on the PULHES categories, and hence recruits with one or more of these scores are typically denied enlistment.

In addition to the PULHES evaluation, each applicant must meet height and body fat standards. Regarding height, applicants must be 60-80 inches in height if male (58-80 inches in height if female).[8] Regarding body fat, each Service begins with sex-specific height/weight charts. Applicants exceeding the maximum weight in the appropriate height/age area of the chart are measured for body fat. Body fat standards vary across Services. The Army's body fat standards vary by sex and age. The body fat limits for enlistment are presented in Table 3.2.

[7] Details of the definitions for each number with regard to each PULHES category appear in Army Regulation 40-501, 1 February 2005.
[8] U.S. Marine Corps applicants must be 58-78 inches tall if male and 58-72 inches tall if female.

Table 3.2. Maximum Percent Body Fat Standards for Enlistment into the U.S. Army

	Age			
Sex	**17-20**	**21-27**	**28-39**	**40 and older**
Male	24	26	28	30
Female	30	32	34	36

Source: Army Regulation 40-501, *Standards of Medical Fitness*, pp. 17-18.

Applicants can receive medical waivers for certain disqualifying conditions. Waiver policy varies across Services, but each Service provides a judgment about each waiver on an individual basis. Therefore, if two individuals with the same disqualifying conditions both apply for a waiver, said waiver may be given to both, only one, or neither, depending on the subjective judgment of the waiver authority. Each Service also has its own waiver authority; for the Army, waiver authority resides with the U.S. Army Recruiting Command (Sackett & Mavor, 2006).

Physical Abilities

Testing of physical abilities varies across the Services. In the Army, applicants typically complete a "1-1-1" in the presence of their recruiter. The "1-1-1" is an assessment of the number of sit-ups and push-ups one can do in one minute, as well as a timed one-mile run. Note, however, that this assessment does not occur at MEPS, as it is an Army recruiting requirement only (Roy Campbell, personal communication, January 15, 2009). The test can occur just about anywhere there is a measured course for the one-mile run. Hence, recruiters might administer the 1-1-1 at a school's track, a gymnasium, or a YMCA. In contrast, the Air Force routinely administers a dynamic lift test

at MEPS to applicants wishing to enlist in particular Air Force specialties.

In 2004, the Accession Medical Standards Analysis and Research Activity (AMSARA) began evaluation of a test of recruit physical ability and motivation that is given at MEPS—the Assessment of Recruit Motivation and Strength (ARMS). ARMS pairs a modified version of the Harvard Step Test (Brouha, Graybiel, & Heath, 1943) with a 60-second pushups test.[9] For the step test, recruits must step up and then back down from a 16" step (for men; 12" step for women) at a pace of 30 steps per minute (with a step comprising of stepping up onto the platform with both feet and then back to the floor with both feet) for 5 minutes or until fatigued to the point that they must stop (recruits must continue for the full 5 minutes to pass the test). To pass the pushups test, men and women must complete 15 and 4 Army-standard pushups, respectively, within the allotted time. The goal of the ARMS evaluation is to determine the impact of enlisting applicants who "pass a fitness test but are disqualified on other grounds" (Sackett & Mavor, 2006, p. 121; see also Krauss, 2004).

Early research (AMSARA, 2005) found that the attrition rate for females who exceeded the body fat limit but passed the ARMS screen did not differ statistically from the attrition rate for females who met the body fat standard. For males, the attrition rate for those who exceeded the body fat limit but passed the ARMS screen was slightly higher than the rate for males who met the standard.

Subsequent research has continued to demonstrate the validity of the ARMS screen. Scott et al. (2008) reported that ARMS, "the

[9] ARMS originally included a dynamic lift test (currently used by the Air Force for assessing qualification for some specialties), but it was removed in January, 2006 to "streamline ARMS testing" (AMSARA, 2005).

first prospective study conducted in the U.S. Army to assess physical fitness prior to accession" (p. 6), provided statistically significant incremental prediction of attrition (beyond other known related variables) in a sample of 7,612 active-duty recruits who accessed between May, 2004 and December, 2005. Specifically, Scott et al. reported hazard ratios of failing ARMS relative to passing ARMS of 1.36 for men and 2.27 for women, indicating that men and women failing ARMS had attrition rates that were 36% and 127% higher, respectively. In this study, no distinction was made with regard to other potential disqualifications (e.g., being overweight). A second study that evaluated the relation between body fat standards and the occurrence of lower extremity stress fractures in females found that female recruits who exceeded the body fat standard experienced a significantly higher risk of experiencing a stress fracture of the lower extremities. No relation between body fat and lower extremity stress fractures was observed, however, for those who passed the ARMS screen (i.e., physically fit females).

The Scott et al. (2008) study involved ARMS screening at six geographically dispersed MEPS. Since then, the Army has increased the scope of ARMS screening to more than 50 locations (El Nasser, 2007).

Moral Character/Conduct

In addition to screening for mental, physical, and medical standing, DoD screens applicants for moral character. Although a rather complex process, screening for moral character and conduct primarily involves reviews of an applicant's history of criminal behavior and drug/alcohol abuse.

Like education status, research has shown moral character status to relate to first-term attrition. Enlistees with moral character waivers attrite at a higher rate than do enlistees without waivers (Fitz &

McDaniel, 1988) and receive a greater number of discharges for unsuitability (Fitz & McDaniel, 1988; Flyer, 1989, 1996; Means, 1983). More generally, enlistees with a record of criminal behavior (whether or not they received a waiver) have higher attrition rates (Putka, Sipes, & Ramsberger, 2002). A recent study (Putka, Noble, Becker, & Ramsberger, 2004) found that recruits receiving moral character waivers continue to attrite for reasons associated with moral character (e.g., misconduct, drug/alcohol abuse) at higher rates than do recruits without such waivers, but found no differences in attrition rates for other types of attrition (e.g., medical).

Moral character screening involves an initial review while visiting with the recruiter and a more thorough investigation by MEPS personnel. Applicants must divulge all arrests or convictions, whether the offenses were committed as a juvenile or had been sealed or expunged. Flyer (1990) noted that more offenders are identified through self-admission than through background checks—perhaps because withholding or falsifying information during this process is a federal offense.

Certain behaviors bar the applicant from enlistment straightaway (e.g., being on probation or parole at application, being under the influence of drugs/alcohol at any point during the enlistment process), whereas other behaviors do not (e.g., one misdemeanor leading to conviction or other "adverse disposition," minor traffic offenses involving fines less than $250). Other behaviors (generally combined with their frequency) can bar applicants from enlistment unless they receive a waiver.

As with medical waivers, the Services consider moral character waivers on an individual basis. They can be granted by the Reception Battalion Commander or the Commanding General

of Recruiting Command. Enlistment waiver policies for moral character differ across Services, although each Service uses a single set of DoD waiver codes to document Service-authorized waivers and follows a similar process (Burnfield, Handy, Sipes, & Laurence, 1999). The Army requires waivers for applicants with records indicating (a) six or more minor traffic offenses (fine of $100 or more per offense), (b) three or more minor non-traffic offenses, (c) two or more misdemeanors, or (d) one or more felonies (http://www.army.com/resources/item/2150).

Other Requirements

DoD has several other miscellaneous requirements that applicants must meet. They are as follows (http://usmilitary.about.com/od/joiningthemilitary/a/enlstandards2_5.htm):

- Citizenship – Applicants who are not U.S. citizens are eligible for enlistment to the military Services (although they cannot be commissioned). Non-citizens must have either a permanent residence visa or green card and must have established a residence and home of record in the United States. Note, however, that non-citizens indigenous to countries that the military considers hostile may not be eligible.
- Dependents – DoD typically bars enlistment to individuals having three or more dependents under 18 years of age. Specific Services have even stricter policies. Applicants to the Army require a waiver if they have two or more dependents (other than a spouse).
- Single Parents – The only option available to single parents is the Army National Guard.

- Spouse of Military Personnel with Dependents – Applicants who have dependents under the age of 18 and are married to a Servicemember require a waiver for enlistment.
- Financial Condition – Some applicants need to demonstrate that they have acceptable financial standing. The Army conducts a Financial Eligibility Determination only in conjunction with evaluation of a waiver for the number of dependents.
- Miscellany – Numerous other specific conditions can require applicants to receive a waiver, if not bar them from enlistment entirely. Conditions include those discharged from service dishonorably or for conscientious objector status and those who have retired after completing 20 or more years of service for the Federal government.

Classification

Classification is the process by which recruits are assigned to their respective military occupations.[10] The assignment can be based on only one or several factors, and these can differ across classification systems, depending on the goal of the classification process. Various factors have been discussed in the scientific literature. Of particular note is the notion of mean predicted performance (MPP) as used in differential assignment theory (DAT; Johnson & Zeidner, 1991; Zeidner & Johnson, 1994). The goal of classification based on MPP is to maximize the predicted levels of job performance resulting from the classification of personnel into available military occupations.

[10] Chapter 5 of this volume contains a detailed discussion of enlisted personnel classification.

Although other maximization/optimization routines and criteria are possible, the military Services currently do little optimal classification. Instead, the primary drivers of classification decisions are recruit qualifications (i.e., does the recruit possess the required test scores to enter a particular occupation) and training seat availability (i.e., which occupations will have spaces to fill in their upcoming training courses). In the Army, career counselors use the Recruiting Quota System (REQUEST) to identify occupations for which recruits have qualifying ASVAB scores (Sticha, Diaz, Greenston, & McWhite, 2007). These occupations are then presented to each recruit in order of the Army's fill requirements.

To supplement REQUEST, the Army has been researching the potential use of the Enlisted Personnel Allocation System (EPAS) for introducing additional optimization goals into the classification process (Sticha et al., 2007). Currently, the design concept involves assigning recruits to occupations such that each recruit will be qualified for his/her assigned occupation and the mean level of predicted performance in each occupation will be maximized. This optimal set of occupational recommendations would then be used to augment the list of occupations provided by REQUEST. As of this writing, EPAS has yet to be used to make operational classification decisions.

Possible Future Enhancements

Various enhancements to the current selection and classification process have been discussed and are being researched. Presented briefly below are four lines of research that could lead to changes in how DoD and the Army determine qualification for entry military service.

Recommendations from the ASVAB Review Panel

In 2006, a panel of testing experts convened to discuss potential improvements to the technical underpinnings of the ASVAB. This ASVAB Review Panel generated 22 recommendations for enhancing the technical quality of the ASVAB testing program (Drasgow, Embretson, Kyllonen, & Schmitt, 2006). Several of these regard the content of the ASVAB, beginning with Recommendation 6: "The contents of ASVAB should be reevaluated" (p. 25). Suggestions for promising additional content appear in several of the other recommendations. For example, the Review Panel recommended development and evaluation of tests of information technology literacy (Recommendation 8)[11] and non-verbal reasoning (Recommendation 9) for potential addition to the ASVAB.

Recommendations 10 through 14 concern noncognitive testing. For example,

Recommendation 10 reads as follows:

> Noncognitive measures should be included in the battery of tests that are used for classification. This will mean that the tests are used primarily to help make a satisfactory job assignment, an objective that the individual and the military share. If this use of the test scores is made clear to applicants, it should diminish their motivation to fake. It may also decrease the ability to fake because there would be multiple different job assignments to consider when responding. (Drasgow et al., 2006, p. 36)

[11] The U.S. Air Force is currently developing an aptitude test of information and communications technology.

The Panel listed four reasons for recommending noncognitive tests: (a) the importance of motivational determinants of performance and service duration (e.g., Ackerman & Beier, 2003; Campbell, McCloy, Oppler, & Sager, 1993) and their demonstrated validity across many occupations (Barrick & Mount, 1991; Hough, 1998; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; Mumford & Stokes, 1992; Reilly & Chao, 1982; Schmidt & Hunter, 1998); (b) their relation to first-term attrition; (c) their potential for incremental validity, given their relatively low correlations with the ASVAB; and (d) the relation of vocational interest measures to job satisfaction, a strong predictor of attrition (R. Hogan & Blake, 1996; Holland, 1997).

The Army has explored non-cognitive predictors for many years now, starting in earnest with Project A and the development of measures of temperament (Assessment of Background and Life Experiences—ABLE), vocational interests (Army Vocational Interest and Career Examination—AVOICE), and work values (Job Orientation Blank—JOB). Army research has seen development of many other potential non-cognitive measures, such as the Work Preferences Assessment[12] (WPA, a measure of vocational interests; Putka & Van Iddekinge, 2007) and the Work Suitability Inventory (WSI, a measure of personality as well as person-organization fit; McCloy & Putka, 2007). The Army has particular interest in three noncognitive measures as of this writing—the Assessment of Individual Motivation (AIM, a derivative of the ABLE), the Rational Biodata Inventory (RBI), and TAPAS.

[12] The WPA was originally named the Work Preferences Survey (WPS).

Assessment of Individual Motivation (AIM)

The Assessment of Individual Motivation (AIM) is a self-report, forced-choice measure of six temperament constructs: Work Orientation, Adjustment, Agreeableness, Dependability, Leadership, and Physical Conditioning. Research has shown the AIM to predict Soldier performance (Horgen et al., 2006; Kilcullen, Mael, Goodwin, & Zazanis, 1999; Kubisiak et al., 2005; White, Hunter, & Young, 2008; White & Young, 2001) and attrition (Hunter, White, & Young, 2008; Young, McCloy, Waters, & White, 2004). AIM is a multidimensional forced choice measure (cf. Jackson, Wrobleski, & Ashton, 2000; McCloy, Heggestad, & Reeve, 2005; Sisson, 1948; Wright & Miederhoff, 1999). Each item comprises four statements, with each statement assessing a different temperament construct. When completing the AIM, respondents must select for each item the statement they believe is most like them, and the one that is least like them (see Figure 3.1).

__(A) I have almost always completed projects on time.
__(B) I have not exercised regularly.
__(C) I have enjoyed coordinating the activities of others.
__(D) I have a hard time feeling relaxed before an important test.

Note. From Young, McCloy, Waters, and White (2004).

Figure 3.1. Sample AIM item.

The Army is currently using the AIM in the Tier Two Attrition Screen (TTAS), an additional screen used to identify the lowest attrition risks among applicants who lack a traditional high school diploma (White et al., 2004; Young & White, 2006).

TTAS serves as one component of a pilot program for expanding the non-high school diploma graduate recruiting market.

Rational Biodata Inventory (RBI)

The Rational Biodata Inventory (RBI) is, as its name implies, an instrument used to determine past behaviors and life experiences of those who complete it, with scoring based on a rational scheme (as opposed to resting on empirical keying). The RBI has taken different forms in prior Army research and operational applications (e.g., for selection into Special Forces) over the past several years. The RBI measures nine dimensions via self-report, Likert-type items. These dimensions—Achievement Orientation, Army Identification, Peer Leadership, Stress Tolerance, Fitness Motivation, Self Efficacy, Traditional Values, Continuance Commitment, and Hostility to Authority—were based on a job analysis meant to identify competencies anticipated to be of importance in the future Army (Sager, Russell, Campbell, & Ford, 2005) and consideration of constructs assessed by other Army biodata measures (Kilcullen, Goodwin, Chen, Wisecarver, & Sanders, 2002; Kilcullen, Mael, Goodwin, & Zazanis, 1999; Kilcullen, McCloy, Putka, & Van Iddekinge, 2005; Kilcullen, White, Sanders, & Hazlett, 2003). The RBI also features a scale for identifying those who distort their responses. The Army is currently conducting research on the utility of the RBI for (a) reducing attrition from the Reserve Officer Training Corps's (ROTC) 4-year scholarship program, where it is called the Officer Background and Experience Form (OBEF; see Putka, 2008); and (b) screening applicants to Officer Candidate School (Kilcullen, Tremble, Babin, Robbins, & Russell, 2008).

Tailored Adaptive Personality Assessment System (TAPAS)

The Tailored Adaptive Personality Assessment System (TAPAS; Drasgow, Stark, & Chernyshenko, 2006; Stark, Drasgow, & Chernyshenko, 2008) is a computerized personality assessment incorporating item response theory to assess 23 hierarchically structured personality facets (22 primary personality facets plus Physical Condition—a facet useful for predicting military attrition but argued to lie outside the traditional latent structure of personality constructs). TAPAS assesses examinees' standings on the facets of interest via multidimensional pairwise preference items (Stark, 2002)—a procedure designed to increase the resistance of the measure to examinee faking. The Army is currently collecting data using the TAPAS-95s, a 95-item measure that assesses 12 of the 23 TAPAS facets (including Physical Condition; see Table 3.3).

Table 3.3. Facets Assessed by the TAPAS-95s

Dominance	Order	Intellectual Efficiency
Excitement Seeking	Non-Delinquency	Curiosity
Cooperation/Trust	Even Tempered	Tolerance
Achievement	Well Being	Physical Conditioning

Summary

The U.S. military attempts to obtain a detailed portrait of the attributes and capabilities of each individual who applies for military service. The present qualification eligibility determination considers myriad factors (cognitive ability, medical/character standing, etc.). Nevertheless, the military continues to research the benefits that might be realized by incorporating other potentially important pieces of information

(e.g., non-verbal reasoning, non-cognitive characteristics, aptitude for information and communications technology) into the qualification evaluation. Such improvements to the enlistment qualification process would increase the probability that future warfighters will be well suited for military service, thereby ensuring that America's military remains the best in the world.

References

Ackerman, P. L., & Beier, M. E. (2003). Intelligence, personality, and interests in the career choice process. *Journal of Career Assessment, 11,* 205-218.

AMSARA (2005). *2005 annual report.* Retrieved December 2, 2008, from http://www.amsara.amedd.army.mil/report/reportindex2005.asp?strReportChoice=3

Association of the U. S. Army (2008, November 1). News call. *ARMY Magazine, 58*(11).

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1-26.

Brose, G. D. (1999). *Could realistic job previews reduce first-term attrition?* Unpublished master's thesis. Monterey, CA: Naval Postgraduate School.

Brouha, L., Graybiel, A., & Heath, C. W. (1943). The step test: A simple method of measuring physical fitness for hard muscular work in adult men. *Review of Canadian Biology, 2,* 86-92.

Buddin, R. J. (2005). *Success of first-term soldiers: The effects of recruiting practices and recruit characteristics.* Santa Monica, CA: Rand Corporation.

Burnfield, J. L., Handy, K., Sipes, D. E., & Laurence, J. H. (1999). *Moral character enlistment standards: Documentation, policy, and procedure review* (FR-WATSD-99-59). Alexandria, VA: Human Resources Research Organization.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35-70). San Francisco: Jossey-Bass.

DiFazio, A. S., McCloy, R. A., Hooper, A. C., & Russell, T. L. (2008). *Interim program evaluation of the Army's Foreign Language Recruiting Initiative (FLRI)* (FR 08-125). Alexandria, VA: Human Resources Research Organization.

DiFazio, A. S., Putka, D. J., Wasko, L., McCloy, R. A., & Shaw, M. (2008). *Final program evaluation of the Army's Foreign Language Recruiting Initiative (FLRI)* (FR 10-15). Alexandria, VA: Human Resources Research Organization.

Drasgow, F., Embretson, S. E., Kyllonen, P. C., & Schmitt, N. (2006). *Technical review of the Armed Services Vocational Aptitude Battery (ASVAB)* (FR-06-25). Alexandria, VA: Human Resources Research Organization.

Drasgow, F., Stark, S., & Chernyshenko, O. S. (November, 2006). *Toward the next generation of personality assessment systems to support personnel selection and classification decisions*. Paper presented at the 48th annual conference of the International Military Testing Association, Kingston, Ontario, Canada.

El Nasser, H. (2007, March 30). Study: Percentage of overweight recruits up [Electronic version]. *Army Times*.

Ellis, H. (1999). *A decomposition analysis of first-term attrition in the U.S. military*. Unpublished master's thesis. Monterey, CA: Naval Postgraduate School.

Elster, R. E., & Flyer, E. S. (1981). *A study of the relationship between educational credentials and military performance criteria*. Monterey, CA: Naval Postgraduate School.

Fitz, C. C., & McDaniel, M. A. (1988). *Moral waivers as predictors of unsuitability attrition in the military* (PERS-TR-88-006). Monterey, CA: Defense Personnel Security Research and Education Center.

Flyer, E. S. (1956). *Factors relating to discharge for unsuitability among 1956 airman accessions to the Air Force*. Lackland, AFB, TX: Personnel Research Laboratory.

Flyer, E. (1989). *Recruits with an enlistment waiver for preservice offenses: Incidence rates, recruit characteristics, and attrition from service*. Monterrey, CA: Defense Manpower Data Center.

Flyer, E. (1990). *Characteristics and behavior of recruits entering military service with an offense history*. Washington, DC: Directorate for Accession Policy, Office of the Assistant Secretary for Defense (Force Management Policy).

Flyer, E. (1996). *California male recruits with a preservice arrest history: Identification, characteristics, and behavior on active duty*. Washington, DC: Directorate for Accession Policy, Office of the Assistant Secretary for Defense (Force Management Policy).

General Accounting Office (1998). *Military attrition: Better data, coupled with policy changes, could help the services reduce early separations*. Washington, DC: Author.

Hogan, P. F., & Harris, D. A. (1994). Policy and management applications of the accession quality cost/performance trade-off model. In B. F. Green, Jr., & A. S. Mavor (Eds.), *Modeling cost and performance for military enlistment* (pp. 129-157). Washington, DC: National Academy Press.

Hogan, R., & Blake, R. J. (1996). Vocational interests: Matching self-concept with the work environment. In K.R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp.89-144). San Francisco, CA: Jossey-Bass.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments (3rd Edition)*. Odessa, FL: PAR.

Hooper, A. C., Paullin, C., Putka, D. J., & Strickland, W. S. (2008). *An empirical analysis of reasons for attrition among first-term airmen in the United States Air Force* (FR-08-32). Alexandria, VA: Human Resources Research Organization.

Horgen, K. E., Kubisiak, U. C., Bruk-Lee, V., Connell, P. W., Penney, L. M., Borman, W. C., Pace, V. L., Lentz, E., White, L. A., Young, M. C., & Bowles, S. (2006). *Evaluation and refinement of a screening instrument for U.S. Army Recruiters: Noncommissioned Officer Leadership Skills Inventory*. (Tech. Rep. No. 1177). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hough, L. M. (1998). Personality at work: Issues and evidence. In M. D. Hakel (Ed.), *Beyond multiple choice*. Hillsdale, NJ: Erlbaum.

Hunter, A. E., White, L. A., & Young, M. C. (2008, August). *Predicting First-Term Enlisted Attrition: The Tier Two Attrition Screen*. Paper presented at the 116th annual meeting of the American Psychological Association, Boston, MA.

Jackson, D. N., Wrobleski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced-choice offer a solution? *Human Performance, 13*, 371-388.

Johnson, C. D., & Zeidner, J. (1991). *The economic benefits of predicting job performance: Vol. 2. Classification efficiency*. New York: Praeger.

Kilcullen, R. N., Goodwin, J, Chen, G., Wisecarver, M., & Sanders, M. (2002). *Identifying agile and versatile officers to serve in the objective force*. Paper presented at the 2002 Army Science Conference, Orlando, FL.

Kilcullen, R. N., Mael, F. A., Goodwin, G. F., & Zazanis, M. M . (1999). Predicting U.S. Army Special Forces field performance. *Human Performance in Extreme Environments, 4(1),* 53-63.

Kilcullen, R. N., Putka, D. J., McCloy, R. A., & Van Iddekinge, C. H. (2005). Development of the Rational Biodata Inventory (RBI). In D. J. Knapp, C. E. Sager, & T. R. Tremble (Eds.), *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria* (Tech. Rep. No. 1168). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Kilcullen, R. N., White, L. A., Sanders, M., & Hazlett, G. (2003). *The Assessment of Right Conduct (ARC) administrator's manual.* (Research Note No. 2003-09). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Kilcullen, R. N., Mael, F. A., Goodwin, G. F., & Zazanis, M. M. (1999). Predicting U.S. Army Special Forces field performance. *Human Performance in Extreme Environments, 4,* 53-63.

Kilcullen, R., Tremble, T., Babin, N., Robbins, J., & Russell, T. (2008). Predicting performance and career intentions among Army OCS candidates. Unpublished manuscript.

Knox, B. W. (1998). *Analysis of the Navy delayed entry program and recruit training center attrition.* Monterey, CA: Naval Postgraduate School.

Krauss, M. R. (2004). *Assessment of recruit motivation and strength (ARMS) phase 2: Preliminary results on weight and disqualifications*. Presentation to the Committee on Youth Population and Military Recruitment: Physical, Medical, and Mental Health Standards, National Academy of Sciences, Washington, DC.

Kubisiak, U. C., Horgen, K. E., Connell, P. W., Lentz, E., Xu, X., Borman, W. C., White, L. A., Young, M. C. (2005). *Concurrent validation of the NLSI for U.S. Army Drill Sergeants* (Study Note No. 2006-01). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Laurence, J. H. (1997). *Does education credential still predict attrition?* Paper presented at the 105th Annual Convention of the American Psychological Association, Chicago, IL.

Laurence, J. H., Ramsberger, P. F., & Arabian, J. (1996). *Education credential tier evaluation.* Alexandria, VA: Human Resources Research Organization.

Lindsley, D. H. (1995). *Recruiting of women.* Presentation at the 1995 Committee on Women in the NATO Forces Conference.

McCloskey, M. A. (1999). *An analysis of the effects of deployment on turnover in the United States Army reserve.* Unpublished master's thesis. Monterey, CA: Naval Postgraduate School.

McCloy, R. A., Harris, D. A., Barnes, J. D., Hogan, P. F., Smith D. A., Clifton, D., & Sola, M. (1992). *Accession quality, job performance, and cost: A cost-performance tradeoff model* (HumRRO Final Report No. FR-PRD-92-11). Alexandria, VA: Human Resources Research Organization.

McCloy, R. A., Heggestad, E. D., & Reeve, C. L. (2005). A silk purse from the sow's ear: Retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods*, *8*(2), 222-248.

McCloy, R. A., & Putka, D. J. (2005). Modeling unit attrition. In W. J. Strickland (Ed.), *A longitudinal study of first term attrition and reenlistment among FY1999 enlisted accessions* (Tech. Rep. No. 1172, pp. 209-315). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

McCloy, R. A., & Putka, D. J. (2007). Work Suitability Inventory. In D. J. Knapp & T. R. Tremble (Eds.), *Concurrent validation of experimental Army enlisted personnel selection and classification measures* (Tech. Rep. No. 1205, pp. 96-116). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*, 730-740.

Means, B. (1983). *Moral standards for military enlistment: Screening procedures and impact* (FR-PRD-83-26). Alexandria, VA: Human Resources Research Organization.

Medsker, G. J., Le, H., & Knapp, D. J. (2005). *U.S. Army Foreign Language Recruiting Initiative (FLRI) final evaluation* (FR-04-52). Alexandria, VA: Human Resources Research Organization.

Mumford, M. D. & Stokes, G. S. (1992). Developmental determinants of individual action: Theory and practice in applying background measures. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3, pp. 61-138). Palo Alto, CA: Consulting Psychologists Press.

Oppler, S. H., McCloy, R. A., & Campbell, J. P. (2001). The prediction of supervisory and leadership performance. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits of personnel selection and classification* (pp. 389-409). Mahwah, NJ: Erlbaum.

Oppler, S. H., McCloy, R. A., Peterson, N. G., Russell, T. L., & Campbell, J. P. (2001). The prediction of multiple components of entry-level performance. In J.P. Campbell & D.J. Knapp (Eds.), *Exploring the limits of personnel selection and classification* (pp. 349-388). Mahwah, NJ: Erlbaum.

Putka, D. (2008). *Initial development and validation of assessments for predicting disenrollment of four-year scholarship recipients from the Reserve Officer Training Corps* (Technical Report). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Putka, D. J., Noble, C. L., Becker, D. E., & Ramsberger, P. F. (2004). *Evaluating moral character waiver policy against Servicemember attrition and in-Service deviance through the first 18 months of service* (FR-03-96). Alexandria, VA: Human Resources Research Organization.

Putka, D. J., Sipes, D. E, & Ramsberger, P. F. (2002). *An evaluation of moral character waiver policy against 3-month attrition* (FR-02-045). Alexandria, VA: Human Resources Research Organization.

Putka, D. J., & Van Iddekinge, C. H. (2007). Work Preferences Survey. In. D. J. Knapp & T. R. Tremble (Eds.), *Concurrent validation of experimental Army enlisted personnel selection and classification measures* (Tech. Rep. No. 1205). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Ramsberger, P. F., Laurence, J. H., McCloy, R. A., & DiFazio, A. S. (1999). *Augmented selection criteria for enlisted personnel*. Alexandria, VA: Human Resources Research Organization.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35*, 1-62.

Sackett, P. R., & Mavor, A. S. (Eds.) (2006). *Assessing fitness for military enlistment: Physical, medical, and mental health standards*. Washington, DC: National Academy Press.

Sager, C. E., Russell, T. L., Campbell, R. C., & Ford, L. A. (2005). *Future soldiers: Analysis of entry-level performance requirements and their predictors* (Tech. Rep. No. 1169). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Scott, C. T., Powers, T. E ., Li, Y., Han, W., Weber, N. S., Gary, J. K., Niebuhr, D. W., & Packnett, E. (2008). *AMSARA: Accession Medical Standards Analysis and Research Activity 2006 annual report*. Silver Spring, MD: Walter Reed Army Institute of Research.

Segall, D. O. (2004). *Development and evaluation of the 1997 ASVAB score scale*. Seaside, CA: Defense Manpower Data Center.

Sisson, D. (1948). Forced-choice: The new Army rating. *Personnel Psychology, 1*, 365-381.

Smith, D. A., & Hogan, P. F. (1994). The accession quality cost/performance trade-off model. In B. F. Green, Jr., & A. S. Mavor (Eds.), *Modeling cost and performance for military enlistment* (pp. 105-128). Washington, DC: National Academy Press.

Stark, S. (2002). A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment [Doctoral Dissertation]. University of Illinois at Urbana-Champaign.

Stark, S., Drasgow, F., & Chernyshenko, O. S. (October, 2008). *Update on Tailored Adaptive Personality Assessment System (TAPAS): The next generation of personality assessment systems to support personnel selection and classification decisions*. Paper presented at the 50th annual conference of the International Military Testing Association, Amsterdam, Netherlands.

Sticha, P.J., Diaz, T. E., Greenston, P. M., & McWhite, P. B. (2007). *Field evaluation of Enlisted Personnel Allocation System (EPAS) enhancements to the Recruit Quota System (REQUEST)* (FR-07-10). Alexandria, VA: Human Resources Research Organization.

Strickland, W. J. (Ed.) (2005). *A longitudinal study of first term attrition and reenlistment among FY1999 enlisted accessions* (Tech. Rep. No. 1172). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

United States (2010). Education enlistment credentials, USAREC regulation 601-101. Fort Knox, KY: Headquarters, United States Army Recruiting Command.

U.S. Department of Defense, United States Military Entrance Processing Command. (1984). *Test manual for the Armed Services Vocational Aptitude Battery*. Chicago, IL: Author.

U.S. Department of Defense, Office of the Assistant Secretary of Defense (Force Management and Personnel). (2007). *Population representation in the Military Services: Fiscal year 2006*. Washington, DC: Author.

U.S. Department of Defense, Office of the Assistant Secretary of Defense (Force Management and Personnel). (1992). *Report to the House Committee on Appropriations: Joint-Service efforts to link military enlistment standards to job performance*. Washington, DC: Author.

Young, M. C., McCloy, R. A., Waters, B. K., & White, L. A. (2004). Introduction: An overview of AIM and the preliminary efforts to support its operational use. In D. J. Knapp, E. D. Heggestad, & M. C. Young(Eds.), *Understanding and improving the Assessment of Individual Motivation (AIM) in the Army's GED Plus Program* (Study Note No. 2004-03). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Welsh, J. R., Jr., Kucinkas, S. K., & Curran, L. T. (1990). *Armed Services Vocational Aptitude Battery (ASVAB): Integrative review of validity studies* (Tech. Rep. No. 90-22). Brooks AFB, TX: U.S. Air Force Human Resources Laboratory.

Wenger, J. W., & Hodari, A. K. (2004). *Final analysis of evaluation of Homeschool and ChalleNGe Program recruits* (CRM D009351.A2/Final). Alexandria, VA: Center for Naval Analyses.

White, L. A., Hunter, A. E., & Young, M. C. (September, 2008). *Update on the Army's Tier Two Attrition Screen (TTAS).* Paper presented at the 50th annual conference of the International Military Testing Association. Amsterdam, The Netherlands.

White, L. A., & Young, M. C. (1998). *Development and validation of the Assessment of Individual Motivation (AIM).* Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.

White, L. A., & Young, M. C. (2001). *Validation of a faking-resistant measure of temperament constructs.* Paper presented at the annual meeting of the Society for Industrial-Organizational Psychology, San Diego, CA.

White, L. A., Young, M. C., Heggestad, E. D., Stark, S., Drasgow, F., & Piskator, G. (2004). *Development of a non-high school diploma graduate pre-enlistment screening model to enhance the Future Force.* Paper presented at the 23rd annual meeting of the Army Science Conference, Orlando, FL.

Wright, S. S.,& Miederhoff, P. A. (1999). Selecting students with personal characteristics relevant to pharmaceutical care. *American Journal of Pharmaceutical Education, 63,* 132-138.

Young, M. C., McCloy, R. A., Waters, B. K., & White, L. A. (2004). Introduction: An overview of AIM and the preliminary efforts to support its operational use. In D. J. Knapp, E. D. Heggestad, & M. C. Young (Eds.), *Understanding and Improving the Assessment of Individual Motivation (AIM) in the Army's GED Plus Program* (Study Note No. 2004-03). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Young, M. C., & White, L. A. (2006, November). *Preliminary operational findings from the Army's Tier Two Attrition Screen (TTAS) measure.* Paper presented at the 25th Army Science Conference, Orlando, FL.

Zeidner, J., & Johnson, C. D. (1994). Is personnel classification a concept whose time has passed? In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 377-410). Hillsdale, NJ: Erlbaum.

PART II

ENLISTED SELECTION AND CLASSIFICATION RESEARCH AND PRACTICE

CHAPTER 3

GROWING AN ARMY

Peter F. Ramsberger
Human Resources Research Organization (HumRRO)

For much of America's history, the terms "selection and classification" were misnomers when applied to military service. Several factors played a role in bringing about this state of affairs. For one, from the times of the earliest settlers, Americans had an ingrained distrust of standing armies. The concept of universal military service was widely accepted as being sufficient when there was a need. However, over the years, experience repeatedly demonstrated that early enthusiasm to take up arms to protect the country was typically followed by reluctance as the conditions of battle and soldiering became apparent. During those times, to the extent that personnel screening occurred, it focused on the physical abilities of the men brought to service. And, in times of true need for military manpower, even the standards that might have been in place were loosely applied. So screening was typically not an issue—finding sufficient men to serve was.

Also throughout much of the nation's history, the variety of tasks soldiers performed was limited. Infantry, artillery, and to a lesser extent, cavalry units, were the core of the Army. Therefore, classifying men into jobs was not the monumental undertaking the military faces today. To the extent that you could, you capitalized on the skills men brought with them to service. Otherwise, assignments were typically driven by where men were needed to support the war effort.

Finally, even the techniques used to screen men in the first World War—rudimentary by today's standards—were just evolving, and so were largely unavailable to those responsible for fielding armies in earlier periods. And, as shall be seen, even the use of relatively simple selection instruments caused much suspicion among Army leaders. When the ranks of soldiers were small, personnel decisions could be based on personal knowledge of each man. Over time, the confluence of the scientific development of methods of selection and classification, the need for large military forces to fight wars on a global basis, and the strict imposition of conscription to obtain those forces, led to the development and acceptance of systematic methods for screening and assigning soldiers.

In the pages that follow, a broad overview of how men were obtained for Army service over the bulk of the country's history is provided. For the aforementioned reasons, there is little said about selection and classification, and even less about research. However, this background helps provide a framework for understanding how the need for military manpower was met over time, and the context in which the pioneers of World War I went about their work.

Before the Revolution[13]

The earliest settlers understood the need for protection in the New World, and brought with them munitions and men who knew how to use them. One of the first acts of the settlers in Jamestown was to construct a fortification for their defense. However, throughout the earliest days of the nascent nation, the fact remained that the resources were not available to fund a permanent Army. This was compounded by the need for able-bodied men to be involved in activities to support and grow the economy. Furthermore, as mentioned, the settlers had an inherent distrust of standing armies, based on the concern that the freedoms they sought in the New World could be wrested away if an Army of trained professionals decided to usurp them. This view would remain dominant through much of this country's history, and would greatly impact the way military-civilian relations were established and regulated.

As a result, the colonies turned to the militia system which had been commonplace in Europe. This at a time when the European nations were beginning to invest in paid, standing

[13] Much of this discussion in this chapter is based on the following sources. Additional sources for specific sections are provided where appropriate. Benn, C. & Marston, D. (2006). *Liberty or death: Wars That Forged a Nation.* London: Osprey Publishing, Ltd. Busch. B. C. (2006). *Bunker Hill to Bastogne: Elite Forces and American Society.* Washington, DC: Potomac Books, Inc. Chambers, J. W. (1987). *To Raise an Army: The Draft Comes to America.* New York: The Free Press. Coakley, R.W. (1969). The winning of independence, 1777-1783. In M. Matloff (Ed.) *American military history.* Washington, DC: U.S. Government Printing Office. Cox, C. (2004). *A Proper Sense of Honor: Service and Sacrifice in George Washington's Army.* Chapel Hill, NC: University of North Carolina Press. Karston, P. (Ed.) (1986). *The Military in America: From the Colonial Era to the Present.* New York: The Free Press. Matloff, M. (Ed.) 1972). *American Military History.* Washington, DC: U.S. Government Printing Office. Weighly, R. F. (1984). *History of the United States Army.* Bloomington, IN: Indiana University Press.

armies. Early on, colonial governments declared that there was a universal requirement for able-bodied, free men of designated certain ages to participate in the defense of their lives, livelihoods, and rights. All of these men were to be armed with personally-owned weapons and participate in regular drills. As the colonies grew in terms of population and territory, militias became more organized. Although there were differences over time and between colonies, the general structure involved the royal governor or the assembly selecting commanding officers. This was often done based largely on the man's popularity, given that the appointee would then be responsible for recruiting soldiers to serve under him. Men would muster periodically to train, but their primary focus remained on practicing the trades that provided for themselves and the local economy. The militia served as a peacekeeping force at home, a recruiting base to form expeditionary units, and a source of munitions.

Over time, relatively permanent units began to form as part of the militia system. As early as the 1670s, volunteers were hired to travel between outposts and warn of imminent Indian attack. As described by Weighley (1984), a two-tier system evolved. The Common Militia was made up of those who served only in response to the expectation that all men would take part. The Volunteer Militia, in contrast, was composed of men who willingly accepted the role of soldier and the service that that entailed.[14]

As was common practice, during the Imperial Wars (1689-1762) regiments were raised for specific campaigns against the

[14] As early as 1634 the Massachusetts Bay Colony sought to establish a network of companies of 30 men each who would be ready to serve at a half hour's notice. This was the forerunner of the Minutemen of the Revolutionary War.

French and their Indian allies. Fighting was typically restricted to the summer months, between planting and harvesting seasons. But recruiting for service became difficult, particularly in areas away from the existing threat. Efforts to secure men frequently turned to the in the bottom rungs of society—landless, rootless individuals whose absence due to militia service would go largely unnoticed. Different colonies employed various strategies to meet their requirements for the effort. In 1709, Connecticut established bounties to lure those in need of a livelihood, eventually moving to a draft that excluded blacks and Native Americans, although they could be hired to serve in place of a man of wealth. In 1754, North Carolina declared militia service to be a useful task for those who might otherwise cause disturbances. In 1759, Massachusetts passed legislation that sanctioned the practice of drafting "strollers."

Overall, the militias served their purpose of warding off enemies while allowing for the nation and its economy to grow. However, there was always a question about the reliability of this system to protect the colonies as a whole. Would there be sufficient forces of the men asked to serve to mount sustained campaigns away from the home fronts? And once recruited, would these men be trained well enough to accomplish their mission?

The Revolutionary War

In May 1775, the 2nd Continental Congress, anticipating possible conflict with England, established the Continental Army. On June 14, 1775, they voted to adopt the New England Army to serve as the Army of the United Colonies. Since that time, this has been recognized as the birthday of the United States Army. On July 2 of that year, George Washington was appointed Commander-in-Chief of the 17,000 soldiers comprising the Continental Army.

With the Declaration of Independence in 1776, the need for more manpower was obvious. In October, Congress voted to establish 88 battalions comprised of 60,000 men who would serve a three-year term. They were to be assembled by individual colonies in proportion to their contribution to the overall population of the country.

Although early enthusiasm for the war effort attracted many to service, it was soon evident that manpower was going to be a major problem. Reports of high mortality rates and poor living conditions among soldiers discouraged many from volunteering, or reenlisting when their terms came to an end. Early on, Washington recognized that this was likely to be a problem. In a letter to Congress in 1776, he said:

> When men are irritated, and the Passions inflamed, they fly hastely and cheerfully to Arms; but after the first emotions are over, to expect, among such People, as compose the bulk of the Army, that they are influenced by any other principles than those of Interest, is to look for what never did, and I fear never will happen; the Congress will deceive themselves therefore if they expect it.

Despite the optimistic plans set forth by Congress, the fighting force never reached their projected troop strength. Moreover, some suggest it was unrealistic to think that so many men could have been taken away from the day-to-day work required to sustain an economy and grow a nation without disastrous consequences. So once again, those who served were largely poor, young whites, blacks, and recent immigrants. As the manpower crisis continued, efforts to "recruit" men included purchasing freedom for indentured servants and pardoning

convicted criminals if they agreed to serve. Bounties of cash and land were offered. In 1777, Maryland, struggling to recruit its quota of soldiers, offered a state bounty of $40 along with a pair of shoes and stockings. Two years later, they upped the ante by including a shirt, hat, and pair of overalls. A study of records of soldiers who came from Virginia found that they were typically 14 to 19 years old, and most were the sons of farmers or farm laborers who served in place of fathers or older brothers. As the war progressed, the need for men led to the drafting of non-citizens, including many captured British and Hessian troops. At its peak strength in October 1778, the Continental Army had 18,000 soldiers in its ranks.

The situation for officers during the Revolutionary War was not nearly as grim. As it was in the British Army, officers were viewed as gentlemen, a necessity if they were to command the respect required of leaders. The distinction between officer and enlisted was reinforced through their standard of living, pay, housing, and dress. The military code of conduct strictly forbade officers from performing duties that were thought best carried out by those of lower station. Although there was an abundance of men who sought appointments, it was difficult to find enough of the proper social standing to assume this role. Congress appointed the generals; political connections were essential to be considered for positions at senior levels. Appointments of field grade officers were accomplished by the colonial assemblies, who were often swayed by the opinions of influential men from the colony. Such conventions succeeded in their goal of clearly delineating the leaders from the led, but they were problematic given the lack of facilities for training those commissioned in the ways of war. Such training was largely carried out by capitalizing on the knowledge of experienced officers, and by studying books devoted to the

subject. These problems led one historian to note that the greatest weakness of the Continental Army during the Revolutionary War period was the lack of experienced officers at every grade (Weighley, 1984).

Given the difficulties in obtaining manpower—always an essential element in successfully prosecuting a war—it seems remarkable that the colonies were able to persevere and establish an independent United States. Although there are disagreements among historians as to what the key factors were that led to this outcome, Coakley (1969) highlights four that are frequently cited: (1) the leadership of George Washington as supported by others in high command, such as Benedict Arnold, Nathanael Greene, and Lafayette; (2) the practical difficulties faced by the British in sustaining a war effort at such a great distance; (3) the dedication of the Continental soldiers, as well as the support provided by the militias, and; (4) the assistance of the French in providing money, supplies, and military support. Whatever the cause, the Americans did prevail and then quickly set about dismantling the very Army that had resulted in victory.

Post-Revolution

Following the Revolutionary War, George Washington felt strongly that the nation should retain a standing Army for its own defense. For the most part, however, his countrymen did not agree. In 1784, Congress reduced the force to 80 men who served chiefly as military police guarding supplies at West Point and other locations. Shortly thereafter the authorized strength was raised to 700, forming 8 infantry and 2 artillery companies whose men would serve a one-year term. Continued troubles with Native Americans on the frontier required this number to be increased at various times over the years. In 1792, Congress

created the Legion of the United States which was led by Major General Anthony Wayne. He was able to train an adept fighting force that saw success on the frontier beyond that which had been witnessed previously. That same year, Congress also passed a new militia act, requiring universal military service in armies assembled by the states. The act authorized the President to call upon these militias in times of need; however no soldier could serve in national service for more than three months in any given year. Upon reaching age 18 all able-bodied, free men were to be enrolled in the militia and, within six months, equip themselves for service with, among other things, "a good musket or firelock, a sufficient bayonet and belt, two spare flints, and a knapsack, a pouch, with a box therein, to contain not less than twenty four cartridges suited to the bore of his musket or firelock."

By 1801, the Regular Army was made up of 248 officers and 3,794 enlisted men in 4 infantry, 2 artillery, and 2 light dragoon companies. The following year, Thomas Jefferson founded the United States Military Academy, which was the first institution of higher education in the country to teach engineering.

At the start of the War of 1812, the Army had grown to some 7,000 men who were widely dispersed protecting the frontier and performing various other duties. As a result, the war effort had to rely heavily on militias, with some 30,000 volunteers called into federal service over the course of the conflict. Fully 86% of those who served were in local units. To strengthen the Regular Army, Congress authorized enlistment bonuses of $40, three months advanced pay, and 160 acres of land. By the end of 1812, some 15,000 men had taken this offer. As the war continued Congress considered conscription for the first time in 1814. When the war came to an end later that year, however, the measure was shelved.

Although the Army retained increased levels of manpower for several years following the war, by 1821 it had been reduced to its pre-war size of around 6,000 soldiers. Weighley summed up the situation by stating that, "Henceforth the American Army would continue to be built upon two foundations, represented now by the Militia Act of 1792 and by the Regulars disciplined under Anthony Wayne."

Mexican War—1846 – 1848[15]

President James Polk's decision to seek war with Mexico in 1846 was controversial at best. The abolitionists claimed that the move to gain new territory in the south was fueled by the desire to extend slavery to a larger portion of the country, and thus make it a more entrenched institution. Despite the rancor over the policy, Congress did authorize the action and war was declared on May 13, 1846. By this time, most northern states had repealed laws regarding compulsory military service, although they were largely retained in the south due to concerns over possible slave uprisings. Given this, and the differences in the views of the campaign in the two areas, it is hardly surprising that most of the over 50,000 volunteers who served in the war were from the south.

Manpower was not a concern at the outset. In 1846, the Regular Army stood at some 8,500 men, many of whom were able soldiers having gained much experience campaigning in the west and serving in ongoing hostilities with the Indians. This number swelled to over 30,000 over the course of the war, many of its new members beginning as volunteers before taking the oath of the Army. Although the war was relatively short, it

[15] Additional source for this section: Foos, P. (2002). *A Short, Offhand, Killing Affair: Soldiers and Social Conflict During the Mexican American War.* Chapel Hill, NC: The University of North Carolina Press.

nonetheless was the case that the initial enthusiasm soon waned as, again, news of harsh living conditions and loss of life reached the home front. By 1847, recruiters were scouring jails and taverns seeking men to "volunteer." Desertions were commonplace. According to Foos (2002), nearly 2,900 Regular Army and 3,900 volunteers deserted over the course of the war, some choosing to fight with the Mexicans. The punishment for deserters who were caught was severe, including branding, flogging, and imprisonment.

In the end, Polk's goals were achieved and the United States gained over 500,000 square miles of new territory. And, once again, as the hostilities came to an end, the Regular Army decreased in size, with some 10,700 soldiers serving by 1849.

The Civil War[16]

The issue of whether slavery would be allowed in America's new territories was among many that eventually led South Carolina to secede from the Union, followed by Mississippi, Florida, Alabama, Georgia, Louisiana, and Texas. Following the seizure of Fort Sumter in Charleston, South Carolina, four more states joined the Confederacy—Tennessee, Arkansas, North Carolina, and Virginia. At that time the Regular Army was made up of approximately 16,000 men, many of whom

[16] Additional sources for this section: Chronology of Conscription, downloaded from http://www.teachervision.fen.com/us-history/resource/5669.html. Logue, L. M. (1996). *To Appomattox and Beyond.* Chicago, IL: Ivan R. Dee. Madden, D. (Ed.) (2000). *Beyond the Battlefield: The Ordinary Live and Extraordinary Times of the Civil War Soldier.* New York: Simon & Schuster. Mitchell, R. (1966). *Civil War Soldiers.* New York: Viking Penguin, Inc. Shannon, F. A. (1928). *The Organization and Administration of the Union Army, 1861 – 1865, Volumes 1 and 2.* Cleveland, OH: The Arthur H. Clark Company. Woodworth, S. E. (Ed.) (2002). *Loyal, True, and Brave: America's Civil War Soldiers.* Wilmington, DE: Scholarly Resources, Inc.

chose to leave to join the Army of the South. Over half of the experienced Cavalrymen opted to fight for the Confederacy, as did nearly one quarter of Army officers.

As with the Mexican War, passions ran high in the early days of the war. Particularly in the south, the pressure put to on men of age to fight came from many quarters. As a result, in the early stages of the war, there was a great rush to service in both the Union and Confederate armies. Similar to the Mexican War, however, reports from the front of high casualty rates and poor living conditions served to dampen enthusiasm over time, and both sides resorted to conscription.

The Union. On April 15, 1861 President Lincoln called for 75,000 militia, and on May 3rd another 60,000 were called for volunteer service. Because there was great confidence that the war would be a short-lived affair, initial enlistment terms were three to nine months. However as the "short war" continued with no end in sight, it became clear that relying on uncoerced volunteers would not provide the manpower needed, particularly given the level of casualties that were being experienced. Therefore, on March 3, 1863, Congress passed the Enrollment Act, which declared that all men aged 20 to 45 were eligible for the draft. A commutation fee of $300 could be paid to avoid service, and men with the means to do so could hire a substitute to serve in their stead. This naturally led to the rise of bounty brokers who would match those who wished to pay someone to serve for them with men seeking large sums of cash. Draft Insurance Societies also sprang up, in which a group of men would pay into a common fund that would be used for the commutation fees of any who were drafted. Draft calls were initiated in July 1863, and then again in March, July, and December of 1864. According to Chambers (1987), 300,000

men were summoned using the draft. Of these, 160,000 were deemed ineligible on physical grounds or because of dependency. Another 52,000 paid the commutation fee, 26,000 provided substitutes, and 40,000 simply ignored the call. The rest were enlisted.

Various changes and additions were incorporated into the draft legislation over time. In 1864, the process of commutation was abandoned. Other measures stipulated that aliens who intended to become citizens would be deprived of their political rights, and possibly deported, if they failed to comply with the conscription measures. Volunteering was encouraged by the institution of enlistment bonuses. Over the course of the war, the federal government paid some $585,000,000 in bonuses, with the states kicking in another $286,000,000. The combined figure was greater than the entire salary of the Army during this period. Between the threat of conscription, the social pressure to enlist, and the enticement of bonuses, very few soldiers were actually drafted into the Union Army. Chambers (1987) estimates that close to 92 percent of the 2,100,000 men who entered service over the course of the war were volunteers.

The draft was far from popular throughout the country. In 1864, "draft riots" broke out in New York City, causing 1.5 million dollars in damage, with anywhere from 24 to 100 persons killed. Many lower income men resisted the draft because they were concerned that, should the Union prevail and the slaves be freed, their jobs would be in danger. However, in the end, 35% of the military-aged population of the north served in the Army during the war, with 40-50% of those in their late teens taking part. Foreign white immigrants made up as much as 25% of the force, with 10% being blacks who were allowed to serve after the 2nd Emancipation Proclamation was signed New Years Day

1863. In return for their service, blacks in the border states were promised their freedom when the war was over.

The Confederacy. In March of 1861, the Confederate Congress put out a call for 100,000 men to volunteer for a one-year term of Army service. The response was overwhelming, but clearly not sufficient once hostilities actually began. Therefore, after Fort Sumter, an additional 400,000 troops were solicited. If the government provided supplies and equipment, the term of service was three years; if you brought your own, it was one year. By August of 1861, there were 200,000 men bearing arms.

The Confederacy faced a crisis of sorts in 1862 when the terms of half of the men who were serving were due to expire. Therefore, in April of that year the Congress passed the Conscription Act, which required all able-bodied white men ages 18 to 35 to serve for three years. It also extended the one-year terms of men currently in service to three years so as to avoid the mass exodus that could have occurred otherwise. The Act allowed those who could afford to do so to pay substitutes to serve for them, a practice that was disallowed through amendments in 1863. As the war progressed and the casualties mounted, other amendments extended the age of service from 17 to 50 years. Exemptions were allowed for certain skilled tradesmen whose efforts were needed to sustain the economy and the war effort. An additional exemption was also added covering men who owned 20 or more slaves. Particularly in the latter case, this caused a great deal of resentment as it was viewed as a means of allowing the wealthy to avoid service. As in the Union Army, desertions were not uncommon among southern troops, with one estimate suggesting that some 104,000 men deserted their units during the war, many returning home out of concern for their families.

In the end, 60% of eligible men served in the Army of the South, with 21% having been drafted. Although the draft was responsible for a greater proportion of the enlistments in the south than the north, in both instances it served as a motivator for volunteers. In the south, with its relatively weaker financial status, bounties could not be relied on as a lure to service. Blacks were a major source of manpower that was ignored until late in the war when offers of post-war freedom were made to bring slaves into the fight. In Alabama and Georgia, there were nearly as many slaves as there were free white citizens. And in Mississippi and South Carolina, there were 100,000 *more* slaves than free whites. Obviously, asking slaves to fight for their continued oppression, and expecting them to do so with any enthusiasm, would have been a stretch. But in the end, the population imbalance between the two sides was undoubtedly one of the factors that affected the eventual outcome.

Over the course of the war, one of four white men of military age in the south died, as did one of 10 from the north. In both cases, a preponderance of the losses was due to causes unrelated to battle, typically infection and other disease. Over 400,000 soldiers were wounded, and it has been estimated that approximately 60,000 amputations were performed. Both sides began the war intending to rely on volunteers for its prosecution. And both sides eventually and reluctantly came to the realization that conscription was a necessity if there was any hope for success. This fact, and the way in which the north and the south carried out their drafts, were to have a great influence on the way in which future American armies were raised in time of war.

Post Civil War

As was true after previous conflicts, the size of the Army decreased greatly following the Civil War. Of course, the

Confederate Army ceased to exist, while the Union Army shrank by nearly half in the year following the war, and by nearly half again in the 10 years after that. Weighley (1984) defines two armies during the period of reconstruction: the Congressional Army whose job was to occupy and keep order in the south, and the Army that was performing its usual duties such as patrolling the Mexican border and controlling the Indians. Chambers (1987) describes a post-war environment that was basically hostile to the military, with northerners concerned about the expenditures involved in maintaining a peacetime Army, and those in the south resentful of the forces that occupied their land. He states that:

> Especially after the conclusion of the Indian Wars on the Great Plains in the 1870s, the Army had little to do except round up renegades from the reservations, maintain a professional leadership cadre, and occasionally help maintain order during labor strikes. While officers were accepted socially, promotion was painfully slow (many served as lieutenants for twenty years) and the enlisted regulars were considered social outcasts. No wonder this period was considered "the dark age" of the U.S. Army. (p. 66)

This period also saw the growth of the National Guard which, in 1891 had over 100,000 officers and men, nearly four times the size of the Regular Army. In that year, states appropriated 3 million dollars to support the formation and maintenance of organized militias. At the same time, the RA cost the federal government $49 million. In the year prior to the Spanish American War the RA stood at just over 2,000 officers and some 25,000 enlisted. According to Weighly (1984), this meant

that the size of the Army was smaller in proportion to the country as a whole prior to entering a war than it had been at any time in its history.

Spanish American War

In 1865 Cubans once again ratcheted up efforts to gain their independence from Spain. As the Spanish employed increasingly harsh measures to calm down the disturbance, President Grover Cleveland, followed by his successor William McKinley, looked on with increasing alarm. When a naval inquiry determined that a mine was responsible for the sinking of the battleship, USS Maine, on February 15, 1898 in Havana harbor, McKinley ordered a blockade of Cuba. Then, on April 11 of that year, McKinley delivered his war message to Congress, citing national interests and humanitarian concerns as principal reasons for American involvement to put an end to the fighting. On April 25, Congress passed a resolution declaring that, starting four days earlier, the United States was officially at war with Spain.

Several days before the official declaration of war, Congress had already acted to strengthen the country's military forces through the Mobilization Act. This authorized a force of some 125,000 volunteers and doubled the size of the Regular Army to nearly 65,000. It also included a provision for special units to be formed, including one to be made up of men thought to have a natural resistance to the types of diseases that could be encountered in a tropical environment. These soldiers became known as the "immunes."

The response to the call to arms was widespread. By May 1898, over 124,000 volunteers had come forward. Another 40,000 answered the call when, in that same month, McKinley asked for an additional 75,000 troops to occupy the Philippines. At the signing of the armistice in August 1898, the Regular Army

stood at some 59,000 men along with 216,000 volunteer troops. This despite the fact that the Regular Army rejected up to three quarters of those who sought to enlist on the grounds that they were unqualified physically, mentally, and/or based on moral standards. Nearly 25 percent of those who sought to join volunteer units were turned away based on results from physicals, and another quarter were dismissed after induction due to inability to meet service requirements. Because of restrictions on National Guard units being assigned outside of the United States, the soldiers in units that reported for duty were enlisted as individuals, with their unit reconstituted following induction. As had been the practice in the militias, National Guard officers were largely selected by the men serving under them. When Congress passed a law mandating that governors select the officers from the units in their state, the move was highly unpopular. Recognizing this, politically savvy governors were sure to engage in consultations to determine who would be popular choices for leadership positions.

The short duration of America's involvement in this war does not reflect its impact on the country. Having successfully liberated Cuba from Spanish dominion, the United States for the first time became an imperial power, with effective control over Puerto Rico, Guam, and the Philippines. This necessarily had an impact on the Army, given that maintaining order in such widespread territories required a larger peacetime military than the country had typically maintained—or wanted—in the past. Although its numbers did decrease somewhat following the war, the Army would never again be reduced to pre-war levels.

Between the Wars—1899-1916

The average size of the Regular Army during the years between the wars was about 80,000 officers and enlisted. Nearly one third of these soldiers were stationed overseas, mostly in the Philippines. Perhaps the most significant developments to effect the Army during this period came in the form of two pieces of Congressional legislation.

The Militia Act of 1903 designated the Guard as the nation's ready reserve.[17] As such, it was now to be under dual state and federal control, which led the way for greater federal involvement in its operation. The Act called for the National Guard to be organized and equipped over a five-year period in a manner that mirrored the Regular Army. Federal funds became available at an increased rate to make this a reality. Mandates were put into place for National Guard training twice a month and for short periods annually. It allowed for the detailing of Regular Army officers to National Guard units, and for joint maneuvers between the two forces. Although the original version of the bill maintained provisions that would make federalization of the National Guard difficult, over time these restrictions were eased to a large extent.

[17] This legislation is known as the Dick Act after its major sponsor Senator Charles Dick, a Major General in the Ohio National Guard

CHAPTER 4

SELECTION AND CLASSIFICATION IN WORLD WAR I

PERSONNEL PSYCHOLOGISTS LEND A HAND

Peter F. Ramsberger
Human Resources Research Organization (HumRRO)

On April 2, 1917, President Woodrow Wilson delivered his war message to Congress. Although he, along with the American people, had been reluctant to get the country involved in the conflict that had been raging in Europe and beyond for some three years, the President explained why this was no longer a viable course:

> On the 3rd of February last I officially laid before you the extraordinary announcement of the Imperial German Government that on and after the 1st day of February it was its purpose to put aside all restraints of law or humanity and use its submarines to sink every vessel that sought to approach either the ports of Great Britain and Ireland or the western coasts of Europe or any of the ports controlled by the enemies of Germany within the Mediterranean.[18]

In fact, in March of that year, three American ships had been sunk by German U-boats, incurring a heavy loss of life. Four

[18] Retrieved from http://www.firstworldwar.com/source/usawardeclaration.htm

days after Wilson's message, on the 6th of April, 1917, Congress passed the war resolution that signaled the United States' entry into the battle. At the time that war was declared, there were some 210,000 men in the Army, including National Guard troops who had been nationalized to serve on the Mexican border. In addition, there were some 97,000 National Guard forces in state service. Recognizing that this was an insufficient force to effectively participate in the war effort, Congress passed the Selective Service Act in May 1917. Unlike in previous periods of conscription, the law included no provision for substitutes or bounties, although local screening boards could grant exemptions. In the first wave of the draft, 7,000 men between the ages of 21 and 30 years old were called up. Over the course of the war, more than 24 million men registered and 3 million were inducted. By the end of the war, the Army's ranks had swelled to over 3.5 million.[19]

At the time of President Wilson's war message to Congress, the Society of Experimentalists was meeting at Harvard. This was a group formed in 1904 by Edward Bradford Titchener, an experimental psychologist, then located at Cornell University. The purpose was to gather the leaders of the major psychological laboratories, along with others they chose to include, to discuss the work in which they were involved. Their final session took place on the day that war was declared, and their attention was drawn to what role psychologists could play in this effort. Robert Means Yerkes, a member of the Society and president of the American Psychological Association, was appointed chairman of a

[19] Karsten (1986) cites figures that he found in a letter written by Newton D. Baker, Secretary of War, to Woodrow Wilson in 1920, in which the Secretary states that some 300,000 men failed to respond to the draft call up and tens of thousands deserted within 30 days of reporting. A reprint in Karsten's volume tells the story of area in rural North Arkansas, where the entire male population failed to serve and the efforts they employed to avoid doing so.

committee "to gather information concerning the possible relations of psychology to military problems."[20] In this initial meeting, they consulted with Captain W. S. Bowen, an instructor in military science at Harvard, and preliminarily decided that an apt subject for their work would be the selection of men to serve in the Army, with a particular emphasis on barring those unfit to serve (i.e., the "feebleminded").

Before beginning this work, Yerkes received a telegram from Dr. George E. Hale who was chairman of the National Research Council (NRC). The two men met and the result was the formation of a psychology committee of the NRC, of which Yerkes became chair. This committee met on April 21, 1917, at the Walton Hotel in Philadelphia. They formed 12 subcommittees to study various means by which psychologists could make a contribution to the war effort. These included committees on (a) the psychological examination of recruits, (b) the selection of men for tasks requiring special aptitude, (c) psychological problems of aviation, (d) problems of incapacity (e.g., shellshock), (e) recreation in the Army and Navy, (f) pedagogical and psychological problems of military training and discipline, and (g) problems of motivation. In the end, these subcommittees were variously successful in regard to their accomplishments, with some failing to meet at all.[21]

[20] Boring, E.G. (1938). The Society of Experimental Psychologists: 1904-1938. *The American Journal of Psychology, 51*. 410-423.

[21] As one example, the Committee on Problems of Emotional Stability, Fear, and Self-Control developed a "Personal Data Questionary." This involved a series of Yes/No questions aimed at detecting psychopathic or neuropathic tendencies. After deriving an initial list of questions, the form was tried out with 1,000 men at Camp Upton, NY and a group of subjects who had been committed on mental grounds. The questionnaire was then pared down to 116 items such as: "Do you feel sad or low-spirited most of the time?" "Do you ever walk in your sleep?" "Are you ever bothered with the feeling that people are reading through your thoughts?" Soldiers were to be assessed through the pattern of their answers, rather than based on any one question. Due to the brevity of America's participation in the war, the form was never actually put into large-scale use.

Certainly one of the most active and influential of the groups was that dealing with testing of recruits.

In regard to testing, some differences surfaced in the views of committee members. In 1908, Henry H. Goddard, a committee member, published an article on the original Binet-Simon scales purported to be measures of intelligence. Yerkes thought there would be great value in developing a similar measure that could be group administered in a short time period. His initial concept was that it would primarily serve the purposes outlined in the group's first meeting—to identify men with cognitive ability so restricted that they could not adequately serve in the Army. However, his view expanded quickly, as he came to conclude that, "We should not work primarily for the exclusion of the intellectual defectives, but rather for the classification of men in order that they may be properly placed in the military service."[22]

Another committee member was Walter Dill Scott who was the Director of the Bureau of Salesmanship Research at the Carnegie Institute of Technology. Before being recruited to come to Carnegie by Walter V. Bingham, who founded the first applied psychology program in the country, Scott had developed a methodology for employee selection. He felt strongly that the greatest service psychologists could perform for the Army would be to develop tests of military-related knowledge that would allow for the optimal classification of men by considering pre-existing knowledge and abilities.

In the end the majority prevailed and elected to follow Yerkes lead and develop a measure of cognitive ability. As we shall see, Scott

However, researchers continued to work on the scale after the armistice, developing several revised versions on which validity studies were conducted (e.g., McCartney & Papurt, 1935).

[22] Kevles, D. J. (1968). Testing the Army's intelligence: Psychologists and the military in World War I. *The Journal of American History, 55*, 565-581.

took a separate path, focusing on classification. Both were huge undertakings accomplished with great efficiency under tremendous pressure. As Yerkes was to note, "Speed counts in a war that costs fifty million dollars a day."

The Army Alpha and Beta

On April 29, 1917, Yerkes presented a plan to Surgeon General William C. Gorgas of the Army Medical Corps entitled *Psychological Examining of Recruits to Eliminate the Mentally Unfit*. This called for the development of a 10-minute test that would serve the purpose of singling out those men in need of further examination to determine their fitness for service. Approval was given, and a group of prestigious psychologists was brought together to develop the test. They included Walter Bingham, Henry Goddard, and Lewis Terman, who was responsible for the original version of the Stanford-Binet intelligence test.

The group met from May 28 to June 9 to develop the initial tests. Even here, some differences arose. While Yerkes' goal was to develop a short test that could be given with maximum efficiency, Terman and Bingham thought it wiser to create a longer instrument that could more accurately identify not only men lacking intellectual capacity, but also those on the opposite end of the spectrum. The candidate list of tests with which they began were evaluated on several criteria, including ease of administration in a group setting; their validity as measures of intelligence; the range of intelligence measured; whether they could be scored objectively and quickly; the degree to which they were susceptible to coaching, malingering, or cheating; how much performance would depend on level of schooling; and how quickly they could be administered. Thirteen kinds of tests were examined seriously, and ten were chosen for trial development and administration (see Table 4.1).

Table 4.1. Initial Army Alpha Tests and Weighting

Test	Contribution to Final Score
Following oral directions	4%
Displaying memory span	5%
Rearranging sentences	8%
Solving arithmetic problems	8%
Answering general information questions	16%
Synonyms and Antonyms	16%
Practical judgment	4%
Number series	6%
Analogies	16%
Number comparisons	16%

The initial trial of the test was conducted by committee members between June 10 and June 23 at several installations around the country. In all, 469 soldiers took both the experimental tests and the Binet scales. The correlation between scores on the new test and Binet was deemed satisfactory at approximately .80. The committee met again from June 25 to July 7 to finalize the instruments.

> At this time the committee had in press five forms of group examination record blanks; an individual record blank, which provides special forms of measurement for illiterates, those who have difficulty with the English language, those who exhibit irregularities suggestive of psychopathic condition, those who are intellectually

subnormal or inferior, and finally, those who are distinctly supernormal; an examiner's guide, which contains directions for the conduct of the examinations; and various types of special record sheet.[23]

In addition to the test materials, the committee developed a plan for the validation of the instruments which was submitted in July. After receiving approval, a second set or trials was run in July and August with 4,000 individuals at four institutions (Brooklyn Navy Yard, Fort Benjamin Harrison, the Regular Army Reorganization Camp in Syracuse, and the National Guard Camp in Nashville). A statistical unit, headed by Edward L. Thorndike, analyzed the resulting data in several ways.

- Mean scores on the test were computed for three separate groups: "adult defectives," regular enlisted men, and students in officer training camps. Results were as expected, with low scores for the first group, average scores for the second, and above average for the last.
- Correlations were computed between scores on the test and officer ratings of the men who took it. These were in the range of .50.
- Correlations were also computed between each of the subtests and the overall score, as well as between the subtests themselves. This provided evidence regarding the contribution of each subtest to the overall score and indications as to whether any of them could be eliminated. All were retained at this stage.

[23] Yerkes, R. M. (1918). Psychology in relation to war. *The Psychological Review, (25)*, 2, 85-115.

Thorndike was also responsible for developing weights for the various subtests to determine their contribution to the overall score. He did so by obtaining the input of other psychologists on how important they felt each was, making estimates of the variabilities of the tests, and roughly estimating their intercorrelations.

Yerkes next developed a plan for an official tryout of the instruments, now known as the Army Alpha. From September to December 1917, the tests were administered to all incoming recruits at Forts Devens, Dix, Lee, and Taylor. Some 65,000 new soldiers were involved. Reviews from Army officials and officers were decidedly mixed. The idea of substituting a psychological test for the judgment of experienced Army personnel did not sit well with many. There was also a good deal of question in some minds as to the importance of intelligence for functioning as a soldier, particularly in certain occupations. Two prominent criteria for assessing the tests in the eyes of the Army were whether their results reflected the opinions of officers who commanded the men, and whether their administration was worth the time and effort involved. Anecdotal evidence suggested that the answer to the first question was "yes," and the committee's emphasis on ease and efficiency of administration ensured that time and effort were held to a minimum. As a result, a proposal to extend testing to all new incoming recruits, both enlisted and officer, was approved on Christmas Eve 1917. Testing began in earnest on January 16, 1918.

Test results came to be translated into grades ranging from "A" for those scoring best to "E" for those doing worst. Statistical tests, similar to those conducted on the first wave of test data were done on the scores obtained in the second test period.

Based on those results, the recommendation was made to delete two of the subtests (memory span and number comparison), modify the weights given to the subtests, and add harder items which would more easily allow for the identification of men of superior intelligence. Testing continued as these changes were implemented and new forms printed.

The plan for testing put forth by the Surgeon General called for a staff of 132 officers, 124 NCOs, and 620 enlisted men to implement the program at 31 camps across the country. An original request to have buildings constructed at these sites specifically for test administration was denied. Instead, Commanding Officers were instructed to find existing space for this purpose, an order that did not sit well in at least some instances. To facilitate the training of those who would be administering the tests, an Army School of Psychology was created at the Medical Officers Training Camp, Fort Oglethorpe, Georgia, in February 1918. Some 100 officers and 300 enlisted men were put through an intensive two-month training course. "The examiners, many of them graduate students in education or psychology, endlessly practiced giving the tests in order to obtain uniformity and accuracy in voice, emphasis, and speed."[24]

In May 1918, testing jumped from 12,000 a month to 200,000. Over the course of the war, some 1,750,000 men took either the Army Alpha or Beta (to be described). The prescribed uses for the test were to identify men with aptitude levels that made them questionable for Army service. Such men were to be interviewed to determine suitability, at which point a recommendation could be made that they be released from

[24] Kevles, D. J. (1968). Testing the Army's intelligence: Psychologists and the military in World War II. *The Journal of American History, 55* (3), 565-581.

service. Additionally, it was suggested that the test results could be used in making assignments so as to avoid having units composed primarily of men of lower or higher abilities. Over the course of the war, some 7,700 men were recommended for discharge based on their results on the Army Alpha/Beta, and another 28,000 recommended for transfer. According to Yerkes (1921) commanding officers came up with other ways to use the test results, including as a guide in selecting men within a company for special duties, as a check of the officer's own estimate of his men, assisting in the selection of men for officer duty, and in forming special companies in which training would either be slowed or accelerated to accompany the abilities of the men assigned to them.

Throughout, there was an uneasy relationship between the psychologists and the professional Army. In 1918, three investigations of the program were ordered. The Adjutant General's Office surveyed Commanding Officers, and received a somewhat tepid response.[25] The Assistant Secretary of War, Benedict Crowell, and the Chief of Staff of the Army, Peyton March, each requested studies involving observations of testing and interviews of men identified as unfit for service. Although the reports resulting from the latter efforts were more positive than the survey results, a General Order was issued on August 14, 1918, that reflected the Army's ambivalence. While it formally established the Psychological Division in the Surgeon General's Office, it also restricted the testing. The program became optional, with Commanding Officers free to use, or not

[25] It should be noted that there were doubts about the tests outside the Army as well. As recounted by Kevles (1968), the noted psychologist Edwin G. Boring commented that the notion that the tests measured innate intelligence was "preposterous."

use, the test results as they saw fit. Further, the role was limited to that of balancing units based on aptitude.

A Note on "Illiterates" and Non-English Speakers

A significant problem encountered by those charged with manning an Army to fight in Europe was the large number of men who reported for duty who lacked basic skills in English, either because they had not been taught to read and write, or because they were non-English speakers. Spring (1972) reports that nearly 30 percent of draftees/applicants during World War I were unable to read and understand newspapers or write letters home. White (1986) cites a "widely publicized statistic" indicating that 24.9 percent of those reporting fell into this category, and suggests that it was "too high." He nonetheless acknowledges the scope of the problem and the large loss of men to the war effort if something wasn't done to address it.

The psychologists who developed the Army Alpha anticipated that a written test of this type would most likely not be suitable for large numbers of those reporting for duty. For that reason, they developed an additional test for those who struggled with reading, including those then called illiterates. The so-called Army Beta was made up of seven tests involving mazes, cube analysis, series completion, digit symbol questions, number checking, picture completion, and geometrical construction. It was carefully administered according to well-rehearsed procedures to men deemed illiterate to identify those with sufficient "innate intelligence" to be of use to the Army.

Among non-English speakers, even those who performed well on the Beta were of little use to the Army if they could not understand the commands given by their officers. To address this problem, development battalions were set up in each National Army, National Guard, and Regular Army divisional

camp in July 1918 (White, 1986). Instruction in English was given for two to three hours daily. YMCA personnel were brought in to supervise these efforts and to be instructors along with soldiers and volunteer civilians. Another course of action which was suggested, but rejected, was to form units of officers and men based on their native language. All such questions became moot when the Armistice was signed in the fall of 1918. However, the lives of more than a few men were undoubtedly positively affected by the development of their language skills in preparation for their service to their country.

The Committee on Classification of Personnel

As mentioned previously, Walter Dill Scott was a participant in the original meetings regarding the contribution psychology could make to the war effort. When his notion for focusing on the assignment of men, as opposed to psychological testing, was overruled at this event, he decided to strike out on his own. Based on work he had done previously, he developed a rating scale for captains. Working through Edward Thorndike, Scott got the attention of Frederick Keppel, Assistant to Secretary of War Newton Baker. Keppel asked Scott to come to Washington to further develop his idea. After making some modifications to the scale, a test was conducted at Fort Myer, Virginia, on July 12, 1917, followed by additional tryouts at the headquarters of the Officer Training Movement, Plattsburg, NY. With this experience, Scott began work on a revised form that was to be used by the Second Officer's Training Camps which were scheduled to begin in September. Scott's scales called for ratings on five dimensions: personal, intelligence, physical, leadership, and general value to service. He met with the Secretary of War to discuss his work, and suggested creating a group that could advise the Army on personnel issues. With permission to begin planning, Scott

developed the notion of creating three groups: a scientific staff, civilian experts to assist in research, and a military panel who would identify problems and help implement solutions. Scott's plan was approved on August 5, 1917, and work began on setting up the Committee on Classification of Personnel.

Scott was made the committee's director, with Walter Bingham serving as executive secretary. Other members included Yerkes and Terman. The committee was staffed with two employment managers and eight university psychologists. At the outset, 16 experienced employment managers were brought in from industry for a 10-week period to get things rolling. Their work was performed under the auspices of the Adjutant General's office.

As described by Lynch (1968), the committee's work was guided by six principles:

1. The principle of functionalization dictated that the control and administration of personnel work should be overseen by a full-time officer and his staff.
2. The principle of human differences noted that there is a great variety and wide range of excellence in human traits.
3. The principle of definite personnel requirements dictated that classification should be guided by detailed statements of the skills and knowledge required in a given job.
4. The principle of organization noted that there is an ongoing need to balance personnel requirements and the supply of available men.

5. The principle of economy of personnel indicated that men should be placed where they will do the greatest good for the service.
6. The principle of morale suggested that, before men are transferred from a unit, the impact of that move on the spirit of the man and the organization be taken into account.

The committee took on several projects to facilitate and implement classification procedures. One involved the adaptation of a document previously developed to assess the occupational qualifications of recruits. Among other things, the "Soldier Qualification Card" sought information on a man's employer, type of work performed in current and previous jobs, education level, military experience, and physical attributes (e.g., height, weight). This enabled personnel screeners to determine if an individual's past experience suggested a particular Army occupation. Similar efforts were carried out for officer personnel. To further fine tune their efforts, 84 "trade tests" were created. When a man claimed skills in an arena, he was administered such a test to gather some notion of his level of expertise. The scoring method developed sought to avoid requiring the person delivering the test to have knowledge of the trade about which they were asking. This was accomplished by having a list of key words that one would use in answering a given question. The examiner simply checked off any of the words mentioned by the examinee. The closer the recruit came to saying all of the words, the higher his score. This information was also entered on the Soldier Qualifications Card. All cards were copied and sent to the central personnel bureau where they were kept on file. When faced with a shortage of men with particular skills, this central registry allowed personnel

specialists to identify whether there were men with the needed abilities available, and where they were located. Over the course of the war, 130,000 men were trade tested, and 973,858 men were selected for assignment to technical units. In all, 450 officers and 7,000 enlisted men were engaged in the classification work.

Other efforts involved conducting detailed job analyses so that civilian-military job equivalencies could be developed. In this manner, specialists were able to determine which jobs required the skills of a machinist, plumber, woodworker, and so on. Such specifications were developed in detail for 600 different trades. In addition, analyses were undertaken to identify the types of skilled and unskilled soldiers which were required in each type of unit. Table 4.2 provides a partial example showing the requirements for a Field Signal Battalion. The first column shows the types of men needed in the Battalion, while the last indicates the civilian occupations personnel specialists should be on the watch for to fill those requirements.

Similar efforts were carried out for officers, with no less than 500 different positions in the Army studied to determine what it is that personnel in each job did. This information was used in developing personnel specifications for use in recruiting, as well as to select men to take on various assignments based on their past experience and training.

Table 4.2. Occupational Needs of a Field Signal Battalion (Source: Strong 1918)

Table of Organization	Skilled	Partly Skilled	Occupation and Code
1 Master Signal Electrician	1		Radio Operator and Constructor (31 w plus wc*) plus engineering knowledge
6 Sergeants 1st Class	5		Radio Operator (31 w)
1 First Sergeant			
5 Chiefs of Sections			
9 Sergeants			Caterer (40 C)
1 Mess Sergeant	1		Merchant or stockkeeper (42 18 s)
1 Supply Sergeant	1		
7 Company Duty	7		Radio Operator
15 Corporals	1	1	12 Radio Operator or Telegrapher (31 w. t)
(1 Company Clerk)			
12 Company	1	12	
Etc.			

* Although not explained in the source, the digits in parentheses represent occupation codes used by specialists.

Committee members also consulted with the General Staff and the Surgeon General's office to plan and implement procedures by which recruits who were found to be lacking in aptitude (or in some other way) could be segregated into development battalions where remediation efforts could increase their chances of serving effectively. On the flip side, the War Service Exchange was created in January 1918 to handle the thousands of written offers to assist the war effort. This involved conducting interviews to determine what role these individuals, many of them with special abilities, could play. In all, some 110,000 written offers of service were received.

The Committee on Classification of Personnel was in existence for 14 months before being integrated with the Central Personnel Branch. The total appropriations for their work amounted to $851,650 (Bingham, 1919).

Conclusion

With the signing of the armistice on November 11, 1918, the work of the psychologists came to an end. By June 30, 1919, more than 2,700,000 soldiers had been discharged. The work which the committee had undertaken was epic in scope given the need to do so much in so little time. They had created, from scratch, the Army Alpha and Beta tests and supervised their administration to 1,750,000 men. Trade tests were created to identify men with skills usable in Army occupations and administered to 130,000 new recruits. Officer evaluation forms were used to select men to serve in this capacity. Throughout the war these evaluations were also used to determine who was performing effectively and who was not, which was done across the board on a quarterly basis. Job analyses were carried out on hundreds of Army positions, on both the enlisted and officer level. And all of this was accomplished during the 578 days that the United States was involved in the war effort.

Although the Alpha and Beta were still used following the war, this practice was soon discontinued. However, testing of recruits did not completely disappear. As Carson (1993) recounts, Army Special Regulation No. 65 put forth in November 1918 set minimum aptitude standards for military entry for the first time. Any recruit whose mental development as measured by the Binet-Simon test was found to be equivalent to a child 8 years or younger was barred from service. New tests were also developed for assessing the abilities of those determined to be "illiterates" and non-English speakers. However, for the most part, the Army

returned to its practice of assessing men based on first-hand knowledge of them.

Despite their short life, the impact of Army Alpha and Beta cannot be ignored. As Kelves (1968) noted, Yerkes was inundated with requests for the test in 1920. Spring (1972) reported that the Alpha was widely used in colleges based on the fact that it had been developed by well-known psychologists, had been administered so widely during the war, and became available at a much lower cost would be incurred by developing tests from scratch. Even so, efforts increased across the country to develop assessments that could be administered in groups for a wide range of purposes, including student and personnel testing. Although this was a development that would likely have occurred without the Army tests, there is little doubt that the visibility of Alpha and Beta hastened the progress of testing in America.

The Alpha and Beta did not escape controversy outside the Army, particularly after the war when there was time for researchers to examine the massive amount of data that had been collected (Samelson, 1977). Several controversial findings emerged from this work, including the determination that the average mental age of those tested was 13 years. This was surprising to many and a much debated result. This led some to challenge the validity of the tests and the way in which the data were being analyzed. Subgroup comparisons were also carried out leading to further disturbing claims. For example, when scores were examined by nationality, the results indicated that men from northern European countries performed better than those from the southern and eastern portions of the continent. This led some analysts to conclude that immigration policies should be formulated to favor individuals from areas where

"innate intelligence" was generally higher. The finding that black men on average scored much lower than whites fed the existing prejudices of many at the time. However, further work showed that the facts were more complicated than they appeared at first blush. Researchers demonstrated that black men from three northern states outscored whites from three southern states. In fact, more and more evidence accumulated pointing to the strong influence of circumstances of upbringing and experience on Alpha and Beta performance. Those wishing to make sweeping statements based on the data collected during the war were eventually dismissed.

Yerkes was clearly disappointed that his efforts and those of his colleagues did not have a more immediate and lasting impact on the military. He authored an article published in 1941 in which he bemoaned the fact that the Germans had accepted military psychology to a much greater degree than had the Americans. After recounting some of their efforts, Yerkes offered the following:

> But unfortunately provisions were not made for the continued development of military psychology as mental engineering within the Army. This chapter of service ended abruptly. There were universal sighs of relief, and the Congress promptly acted on the assumption that the war to end wars had actually accomplished that purpose. Isolation had its field day, and instead of capitalizing our precious experience and consolidating the encouraging progress which had been made in mental engineering, we virtually ignored the opportunity in favor of our commercial pursuits. For twenty years military psychology remained for us merely a matter of historical and academic interest.

Although it may be true that concentrated efforts at applying psychological principles for improvement of military functioning were largely absent between the wars, the efforts of the Experimentalists during World War I certainly had an impact as America prepared for the next global conflict.

Sample Items—Army Alpha

Math items

Mike had 11 cigars. He bought 3 more, and then smoked 8. How many cigars did he have left?

A U-boat goes 6 miles an hour under water and 20 miles an hour on the surface. How long will it take to cross a 100-mile channel if it has to go three-fifths of the way under water?

Practical Judgment

If a drunken man is quarrelsome and insists on fighting you, it is usually better to

- knock him down
- call the police
- leave him alone

Aeroplanes failed for many years because

- they were too heavy
- the materials cost too much
- the motor was not perfected

Disarranged Sentences (arrange sentence correctly and then indicate whether it is true or false)

wood eat and good to are coal

seldom forever good lasts luck

Number Series Completion (indicate next two numbers in series)

5 9 13 17 21 25

3 6 8 16 18 36

Analogies

egg—bird :: seed—grow plant crack germinate

Caucasian—English :: Mongolian—Chinese Indian negro yellow

Information

Becky Sharp appears in Vanity Fair Romola The Christmas Carol Henry IV

General Lee surrendered at Appomattox in 1812 1865 1886 1832

Figure 4.1 Sample Army Alpha and Beta items

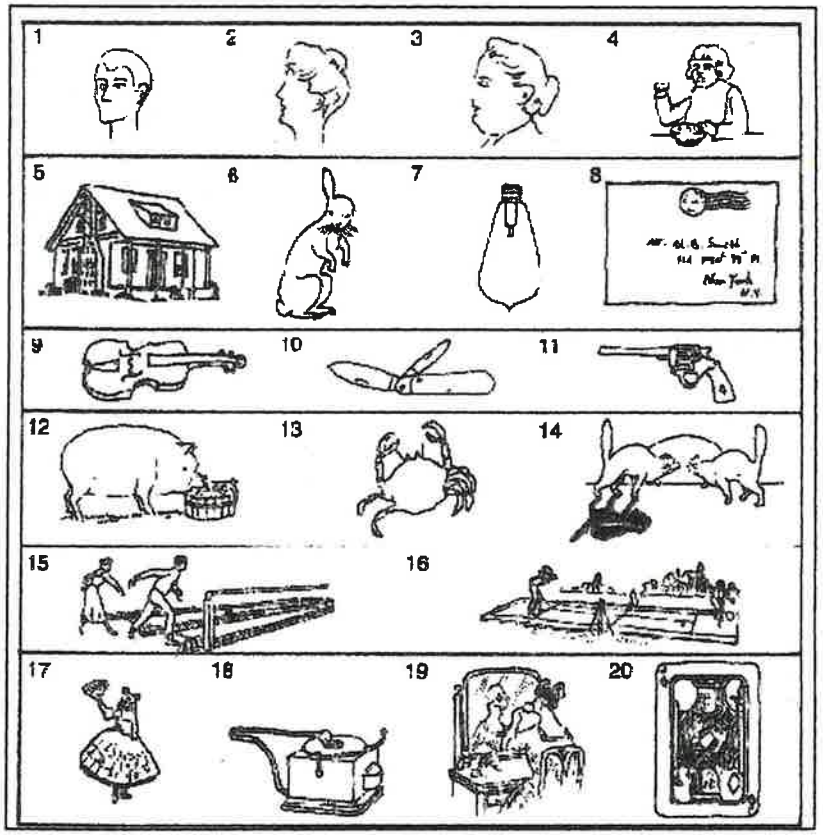


Figure 4.1. Sample Army Alpha and Beta items (Continued)

References

Bingham, W. V. (1919). Army personnel work: With some implications for education and industry. *Journal of Applied Psychology*, *3*, 1-12.

Boring, E. G. (1938). The Society of Experimental Psychologists: 1904-1938. *The American Journal of Psychology*, *51*. 410-423.

Carson, J. (1993). Army Alpha, Army brass, and the search for Army intelligence. *Isis*, *84*, 278-309.

Karsten, P. (Ed.) (1986). *The military in America: From the colonial era to the present*. New York: The Free Press.

Kelves, D. J. (1968). Testing the Army's intelligence: Psychologists and the military in World War I. *The Journal of American History*, *55*, 565-581.

Lynch, E. C. (1968). Walter Dill Scott: Pioneer industrial psychologist. *The Business History Review*, *42*, 149-170.

McCartney, J. L., & Papurt, M. J. (1935). Preliminary report on an objective personnel data questionnaire. *Journal of Abnormal Psychology*, *30*, 95-106.

Samelson, F. (1977). World War I intelligence testing and the development of psychology. *Journal of the History of the Behavioral Sciences*, *13*, 274-282.

Spring, J. H. (1972). Psychologists and the war: The meaning of intelligence in the Alpha and Beta tests. *History of Education Quarterly*, *12*, 3-15.

Strong, E. K. (1918). Work of the Committee on Classification of Personnel in the Army. *Journal of Applied Psychology*, *2*, 130-139.

White, B. (1986). The American military and the melting pot of World War I. In P. Karsten, (Ed.) *The military in America: From the colonial era to the present* (pp. 301-312). New York: The Free Press.

Yerkes, R. M. (1918). Psychology in relation to the war. *Psychological Review*, *25*, 85-115.

Yerkes, R. M. (1921). *Psychological examining in the United States Army. Memoirs of the National Academy of Sciences*. Washington, DC: U.S. Government Printing Office.

Yerkes, R. M. (1941). Psychology and defense. *Proceedings of the American Philosophical Society*, *84*, 527-542.

CHAPTER 5

WORLD WAR II—BUILDING ON EXPERIENCE

Peter F. Ramsberger
Human Resources Research Organization (HumRRO)

The Lead Up to War

Some twelve years after the end of the war to end all wars, international tensions began to surface once more. In 1931, Japan invaded Chinese Manchuria, occupying the whole province the following year, which led to the start of the second Sino-Japanese war in 1937. The Spanish Civil War raged from 1936 to 1939, drawing the involvement of both Hitler's Germany and Mussolini's Italy in support of Francisco Franco. Hitler's expansionist desires became clear with the occupation of Austria, Czechoslovakia, and eventually, in 1939, Poland. The latter intervention forced the hand of both Britain and France, who declared war on Germany in response. There was no immediate military action, however, and the so-called "phony war" lasted until May 10 of the following year when the Germans invaded France. At that point, the seemingly inevitable hostilities were underway.

As it was before World War I, there was little sentiment within the United States in favor of involvement in these conflicts. The memories of 116,708 dead and 205,690 wounded in the first World War were still relatively fresh in the minds of many. In the face of strong isolationist views among a wide spectrum of the American public, President Roosevelt was limited to simply

providing munitions to the western European democracies. However, the Japanese attack on Pearl Harbor on December 7, 1941 ended the illusion that America could remain an onlooker in the hostilities that were quickly engulfing the world. The next day, Roosevelt addressed the Congress, who quickly responded with a formal declaration of war against both Germany and Japan.

Despite the reluctance on the part of much of the country to become involved in the wars abroad, early steps were taken to prepare should this become unavoidable. On September 8, 1939, the president issued an emergency order to increase the size of the Regular Army (RA) and National Guard (NG). At that time, the RA stood at some 190,000 enlisted and officer personnel. In the spring of that same year, W.S. Hunter, a member of the National Research Council (NRC), approached the Army Adjutant General, Major General Emory S. Adams, and offered the services of psychologists in the NRC with personnel issues. His offer was politely declined.[26] Undaunted, Hunter and his colleagues at the NRC nevertheless established the Emergency Committee on Psychology to be prepared to offer the Army expert advice on such matters as identifying men unfit for service and classifying those who passed the screens.

The Adjutant General was not oblivious to the Army's need for action in the area of testing. In the spring of 1939, about the time

[26] Napoli (1978) reports even earlier efforts on the part of behavioral scientists to assist during the lead up to war. In September 1938, Walter Dill Scott, who played such a major role in the personnel classification efforts during World War I, offered to help in a similar manner, but was turned away by the War Department. Two months later, Horace B. English who was executive secretary of the American Association for Applied Psychology contacted Major General Adams to suggest that the time was right to begin work if testing was to be part of the Army's enlistment and classification program. After some back and forth, English was allowed to make a presentation to the Adjutant General's staff, which opened the door for the efforts of others.

he was refusing help from the NRC, he established a Personnel Research Section (PRS) in his office to work on creating a replacement for the Army Alpha and Beta. After about one year, it became apparent that expert outside advice would be of value. On April 2, 1940 Adams asked the NRC to establish the Committee on Classification of Military Personnel. Walter V. Bingham, who had played such a significant role in the personnel work done in conjunction with the first World War, was named chair. Throughout the war, the committee served in an advisory capacity to the PRS in the wide variety of efforts which they undertook.

The first priority of the PRS was to develop a test that would allow for the broad classification of men based on their ability to absorb military training. Other concerns included creating methods for selecting and evaluating officers, updating and improving the occupational interviews and trade tests used in the first World War, and developing supplementary tests to identify men with aptitudes in specialized areas even if they had little or no relevant prior experience. The PRS started with only a small staff, but eventually grew to include 18 personnel technicians, 8 Army officers, and 50 clerks in supportive roles. Various subsections were formed within the office to deal with specific tasks or issues, including occupational analysis, test scoring and records maintenance, and statistical analysis. It was fully understood that the PRS existed to serve the needs of the War Department—research for the sake of research was unacceptable.

Work on initial test construction continued throughout 1940, as did the process of mobilizing for war. On August 27 of that year, Congress passed a joint resolution authorizing the president to call up the NG. Less than a month later, with little pressure from Roosevelt, Congress enacted the Selective Training and Service Act. In its initial form, this act required all

males ages 21 to 35 (inclusive) to register for the draft. Men would be selected through a lottery system and, if called, would be required to serve for 12 months. To appease those who passionately felt the United States should remain neutral in the war, a provision was included in the bill stipulating that men who were drafted would serve only in the Western Hemisphere or in United States possessions in other parts of the world.[27] By the time the act was passed by Congress and signed by the president, the PRS was in the process of validating their newly-created test, and by January 1941 all men who had been in the RA or came in through the NG had been classified on the basis of the new exams and related procedures.

The Army General Classification Test

At the first meeting of the Committee on Classification of Military Personnel, the work-to-date of the PRS was outlined. Colonel M. W. Richardson described the specifications that guided test development, including that the test:

- include both verbal and non-verbal items
- emphasize items calling for spatial thinking and quantitative reasoning
- keep to a minimum items that rely on education (e.g., unlike the Army Alpha, no information-type items)
- minimize the importance of time or speed so as to not penalize able men who simply work at a slower pace

[27] Amendments to the Act were subsequently approved on December 13, 1941 that lifted the territorial restrictions on service and extended the period of service to six months following the conclusion of any war in which the country was involved.

- not serve as a test of trade knowledge
- not seek to measure personality traits
- appeal to the average prospective enlisted man or officer as a reasonable instrument for placing and classifying soldiers

Other concerns were that the test include both easy and difficult items and that it be readily scored by hand or machine.

Richardson also discussed the intended method of reporting results. Unlike the Alpha, there would be no attempt made to establish age-equivalents for the Army General Classification Test and, in fact, such efforts on the part of others would be discouraged. Instead, the test would be scaled in units of standard deviation based on the average performance of a representative population of military-aged men. The test was to include practice exercises and be constructed in a spiral-omnibus form, in which successive blocks of items of each type included in the test are presented (e.g., vocabulary, math). The committee also decided that only one score should be derived based on test performance because of the lack of reliability often found in partial scores. Finally, great care was taken to avoid calling the measure an intelligence test. Harrell (1992) stressed this fact, noting that reports of low "mental age" among World War I soldiers led the public to conclude that "either draftees were dumb, or psychologists were crazy, or both" (p. 876).

By June of 1940, some 5,000 test items had been constructed, including arithmetic reasoning, common sense, vocabulary, number series, and synonyms/antonyms. In an effort to stay true to the mandate of including nonverbal questions, various types of "pictorial" items were developed using cubes, plane figures,

and the like. However, an experimental test demonstrated that there was unacceptable restriction in the range of item difficulty, and different types of item combinations were considered.

The first version of the Army General Classification Test (AGCT-1a) included 150 vocabulary, arithmetic, and block counting items to be completed in 40 minutes. AGCT scores were reported in the same way that the Armed Services Vocational Aptitude Battery scores are presented today; that is, five categories ranging from I representing the highest-level performers to V which includes the lowest scoring. The AGCT was administered to a sample of 3,790 RA enlisted men along with 606 Civilian Conservation Corps enrollees. After eliminating those who were not in the 20-29 year age range, the final sample on which statistics were calculated was 2,675. Five steps were then undertaken to develop an Army Standard Score:

1. Census data from 1930 were used to estimate the proportion of the total Army population in each age group by education and geographic area of residence.
2. These same proportions were then calculated for the sample based on data collected from its members.
3. The sample data were then weighted to reflect the estimated population.
4. Population estimates were derived using the weighted raw score sample means and standard deviations.
5. Linear transformation was used to calculate Army Standard Scores equivalents for each AGCT raw score.

This technique involved several assumptions which, as it turned out, were not quite met. For instance, to obtain accurate outcomes it would have to be the case that age, education, and

geographic location would account for the bulk of the variance in test performance, and that the Army population would be a random sample of the population at large. When researchers compared the percentage of sample members in each category (I – V) with that which would be expected if the scores were normally distributed, some anomalies were discovered. Lower percentages of sample members scored in the Category III (90-109) and IV (70-89) range than expected, with higher percentages in the other groups, particularly Category II (110-129). They pointed to a number of potential causes of this outcome, including the fact that they were unable to include race as a weighting variable in determining Standard Scores, and that the sample was somewhat skewed because men with much lower mental ability and those with deferments were not included. Based on PRS reports, it appears that they decided this problem would not affect the functional utility of the test and set about designing alternate forms.

AGCT-1b included the same block counting items as the first form. The chief difference between the two tests was in the way the vocabulary items were presented. In AGCT-1a, a word was presented with four alternatives and the examinee had to pick the closest equivalent. In the new version, the target word was highlighted in a sentence (i.e., in context), and the examinee had to pick from the four alternatives. AGCT-1b was standardized on a population of 3,856 men who were given both tests. Raw score distributions were calculated for 1-a and 1-b, and simple linear equating methods were used to place the 1-b raw scores on the 1-a score scale, so that the transformed standard scores for 1-b would have the same mean and standard deviation as the standard scores from 1-a. In the end, the score distribution was similar to the first form.

Changes to AGCT 1c and 1d included reducing the number of practice items and presenting them in the test booklet itself rather than as a separate document. In addition, the overall number of test questions was reduced from 150 to 140. The block counting items were retained from earlier versions. A sample of 1,782 soldiers was administered both forms in counterbalanced order, and AGCT-1a scores were obtained from records. Statistical analyses indicated that these new forms were somewhat more difficult than their predecessors, with 2-3% fewer men placed in the upper grades, and corresponding percentages scoring in the lower categories.

The final version of the AGCT used during the war was the AGCT-3a. This was composed of four separately timed and scored tests, including reading and vocabulary, arithmetic comprehension, arithmetic reasoning, and pattern and analysis. Research demonstrated that AGCT-3a yielded scores that were practically identical to the first version of the test.

One administrative change that occurred during this period was that the bottom level score for Category IV was lowered. This appears to have been done, not based on performance data or sophisticated analytic results, but rather because complaints were being received that too many Category V men were being sent to units. By simply broadening the score range for Category IV, and thus reducing the range for the lowest Category, this problem was solved.

Various statistics were calculated to assess the validity of the AGCT. For instance, test results were found to correlate highly with other similar measures such as the Army Alpha (.79), the Otis Higher Mental Ability Examination (.83), and the Army-Navy College Qualification Test (.75). Correlations were also run between AGCT scores and various training performance

scores, with more mixed results. Researchers found that AGCT was moderately related to scores in clerical (.33 to .62), mechanical (.32 to .69), and radio (.24 to .49) classes, with generally lower correlations in driver (.13 to .31), officer candidate (.09 to .46), and Military Academy (.12 to .43) outcomes. The test was not found to be correlated with age or leadership ratings, but was highly related to education (correlation of .73 with highest grade completed and .66 with number of years in school).

Between 1941 and 1946, the AGCT was administered to some 12,000,000 men and women in the combined forces. It was of far greater practical significance than the Army Alpha, both in terms of the number of individuals who took the exam and the importance it had in the selection and classification process. Unlike the Alpha, which was not ready for administration until months after the country began the process of enlisting men to go to war, the AGCT was available from the start of conscription, and thus played a role in the decision making regarding whether and how to use nearly all of the personnel who were drafted or volunteered to serve.

Officers

With the growth of the Army enlisted force from some 250,000 in 1940 to over 7,000,000 five years later, an obvious concern was with finding the men who would be responsible for leading the troops. In 1940 there were just over 18,000 officers in the RA. Bingham (1942) reports that, in the early years of the war, the need for officers was largely met through the pool of 124,000 in the Reserves. By the end of 1943 there were approximately 19,000 NG officers in federal service. Over the course of the war, 180,000 were drawn from the Officer

Reserve Corps, with another 300,000 trained in newly-formed Officer Candidate Schools (OCS).

Enlisted men were selected to apply to OCS based on a variety of factors, including performance on the AGCT and other selection tests, background and experience, and the evaluations of their company officers during the first few months of service. Typically, five months of service was required before consideration would be given. Boards charged with making selections for OCS were provided the *Soldier's Qualification Card* which summarized test scores, past experience, education, and other relevant information. They also received reports from commanders and supplementary data. Each candidate was then interviewed before a final determination was made as to whether OCS was in his future. In some cases, applicants were required to take "searching examinations" to assess their ability to deal with new problems and to master educational content that was essential in their particular field.

Another source of officer personnel was direct commissions for men with specialized civilian training needed by the Army. Approximately 100,000 officers were awarded direct commissions, with less than half being doctors, dentists, or chaplains. Military training was provided before duty assignments were assumed. Early in the war the Volunteer Officer Candidate (VOC) program was also used to recruit officers. This involved tapping into the large pool of men who had been deferred from military service, most often for reasons of dependency. The understanding was that if a given volunteer was not selected as an officer while undergoing basic training, or if he failed to complete OCS, he could return home and assume his former draft status. By the close of 1943, just over 38,000 officers had been commissioned through this program,

and another 27,000 were in OCS. The VOC was largely abandoned in 1943 when there was no longer a shortage of officers.

The Army Specialized Training Program (ASTP) was initiated in December of 1942. Under the ASTP, selected soldiers who had completed basic training were sent to colleges and universities around the country to continue their education, along with further military instruction. The program arose for a variety of reasons, including the recognition that the educational facilities of the Army would be insufficient given the need for large numbers of officers and enlisted men. Another positive outcome of the program was the benefit it would provide to post-secondary educational institutions in the country, which would be facing hard times in the face of a large-scale draft of men as young as 18.

Although it was assumed that most of the ASTP participants would, in fact, become officers, service in other specialist roles such as engineers, medical personnel, and linguists was also envisioned. The maximum number of men allowed in the program at any one time was 150,000. Further, it specifically targeted enlisted men under the age of 22 who had an AGCT score of at least 110. Various other tests were used over time to select men for the program, including the *Army-Navy Qualifying Examinations* constructed by the College Entrance Board, a mathematics inventory used to channel soldiers/students into the right curriculum at the right level, and a series of tests for professional medical training. In addition, more than 150 different national achievement tests were developed to assess program success as demonstrated by student knowledge. The ASTP was phased out in 1943 and 1944 due to Soldier shortages which dictated that men were

needed in the Army and not in college. Perhaps the biggest legacy of the ASTP, however, was the introduction of thousands of men to a college environment which many may never have experienced, and to which many would return after the war. A survey of some 8,000 enlisted men assigned to engineering courses through the ASTP found that 30 percent of respondents indicated that they had received more education because of the program than they had planned on seeking prior to the war.[28]

The PRS also developed tests, rating forms, officer evaluation reports, and interviewing procedures to assess background characteristics, learning ability, and leadership qualities of officers and officer candidates. For instance, the *Higher Examinations (H-1 and H-2)* were specifically created to more finely discriminate between men at the higher levels of AGCT. However after research demonstrated that speed played too much of a role in performance, giving an undue advantage to younger prospects, the *Higher Examinations* were discontinued. Several iterations of *Officer Candidate Tests* were developed, and the final forms, OCT-1 and OCT-2, were found to be highly correlated with AGCT and years of education. However, both tests were found to have higher validity in predicting academic success in OCS than did the AGCT. Various studies were also performed to find methods of assessing leadership potential. Projective instruments such as the *Rorschach* and *Thematic-Apperception Tests* were found to have low correlations with leadership ratings as well as being impractical to administer on a large scale. Other attempts to assess leadership also had disappointing results and were difficult to carry out because of

[28] Among those who took part in the ASTP were former Secretary of State Henry Kissinger, former mayor of New York Ed Koch, newscaster Roger Mudd, sports commentator Heywood Hale Broun, author Gore Vidal, and actor/writer Mel Brooks.

the demands of the war. Finally, a series of tests and associated instruments were developed for the purpose of selecting from the pool of temporary officers those who, if they desired, should be retained in the Army following the war. These included an *Officer Classification Test*, a test of general educational achievement, a *Biographical Information Blank*, and an *Officer Evaluation Report*.

Other Tests for Classification

In addition to the AGCT, the PRS developed a variety of other tests for the purpose of properly placing soldiers in occupations in which they had a reasonable chance of success. As in World War I, trade knowledge tests were created to be administered to men who indicated that they had experience in various fields. The PRS developed numerous editions of tests in electricity, radio, and automotive mechanics to assess knowledge in each of these areas. Informal information tests were also produced in these areas along with driver information for use at training centers to select men for specialized training. Other aptitude tests included:

- *A Meteorology Aptitude Test* was made up of a mental alertness measure, fifty meteorology items, and 144 true-false physics questions
- An *Aircraft Warning Aptitude Test* which involved locating points on a map by grid coordinates and plotting coordinates. A follow-up *Aircraft Warning Classification Test* was given to classify men into specialties. In combination, these tests were said to eliminate 90 percent of failures.

- Selection tests were developed for a variety of other specialties including Balloon Barrage, Military Police, and Medical Technician courses.
- A battery of tests was developed to screen for soldiers who could be successful in Combat Intelligence training, including identification of aerial photographs, route tracing, perception of detail, map reading, and map orientation.
- A variety of measures were evaluated for selecting soldiers to enter the Air Corps basic training center, with two finally being identified as having the highest validity—the *Nut and Bolt Manual Dexterity Test* and the *U-Bolt Assembly Test*.
- Methods to measure truck driving ability included actual road tests with ride-along evaluators checking objective scoring sheets, driver information tests, assessments of visual acuity, and measures of sensori-motor coordination. Some 40 studies were carried out to determine the most effective and efficient means for selecting truck drivers. As of 1943, reports by the PRS staff suggested that driver information tests and experience inventories showed some promise in experimental use, but that the data were incomplete. Methods such as road tests were also found to be difficult to implement on a large scale.
- A *Military Knowledge Test* using a multiple-choice format in pictorial form was developed to assess the level of basic military knowledge required of all soldiers as a means of determining if those who were being redeployed needed remedial training.

- *Educational Achievement Examinations* were created for a wide range of subjects, including algebra, German, physics, plane and solid geometry, and U.S. history.

Perhaps the most widely used tests developed by the PRS were those used to assess mechanical and clerical aptitude and to select soldiers to become Radiotelegraph operators. The first version of the *Mechanical Aptitude Test* was released in February of 1941, followed by two more forms the following October. These included items on mechanical movements, surface development, and shop mathematics. The first two versions were found to be good predictors of performance in mechanics courses, but they were abandoned with the release of AGCT-3a, which included a surface development section. The *Clerical Aptitude Test* was completed in 1940 and included 280 items on name checking, coding, catalog numbers, verbal reasoning, number checking, and vocabulary. Despite the effort to develop the test, it was found to be lacking in predicting clerical grades, which the AGCT was better able to accomplish.

The selection test for radio operators was of critical importance due to the nature of the job, its importance to the Army (nearly five percent of men entering the Army became radio operators), and the rarity with which men entered service with the ability to perform this job. In addition to the skill involved in coding and decoding text, radio operators were required to be adept in the use of other forms of communications (e.g., visual signaling, flags, pyrotechnics), typing, packing and moving equipment, reading maps and aerial photographs, and the use of elementary cryptography. Experience demonstrated that many men failed training in this field, most because they were unable to master the skill involved in coding.

The *Radio Operator Aptitude Test* (ROA) was developed before the war and, at least during some periods, was given to all men who scored 80 or above on the AGCT. If many AGCT high scorers were available, those in the upper ranges were selected for screening. Basically, the ROA involved the presentation of a series of sound patterns ranging from 2 to 13. The sounds were administered by phonograph record, and the task was to identify contiguous patterns as being the same or different. It took 9 minutes to present the entire series of 78 pairs of code patterns. Because of problems with low reliability, it became standard practice to administer it twice to each group of men with a two-minute interval between repetitions. Research demonstrated that the test had considerably higher validity with those who had some experience in this type of task than it did with neophytes. Therefore, PRS personnel developed a *Code Learning Test* (CLT), in which the codes for six punctuation marks were taught and reinforced in a 30 minute period. The test consisted of 100 items in which the six learned items were presented in random order with 28 that were not taught. Examinees were instructed to enter a "0" on his answer sheet if the code was unfamiliar, and record the character for those recognized. Two variations of the CLT were also developed—the Substitution and Code Rhythm tests, however the first instrument was found to be most valid in screening for ability to succeed in training.

Research with the CLT showed low relationships to musical experience, which countered the results found in other work done in this area, and little correlation with preference for receiving radio operator training. Researchers also found little relationship between level of education and ability to learn code, and a low but positive relationship to AGCT score. The CLT was found to have higher reliabilities than the ROA, however over the course of the

war both were used to identify candidates for radio operator training.

The Screening Process

As described by Bingham (1942), the screening process for soldiers began at the induction station, where two primary criteria were examined. The first was literacy. When it was suspected that this might be lacking in a registrant, he was given a short, self-administered *Minimum Literacy Test* which served to identify those whose reading skills were below the fourth grade level. For those who failed, this was followed by a *Visual Classification Test* which was intended to identify those men with sufficient mental capacity to absorb Army training. A 1943 article by the PRS staff indicates that, at that time, no more than ten percent of white and ten percent of black men inducted in a given day could be "illiterates".[29] A more detailed discussion of the literacy problem is provided below.

The second of the first two screens was physical condition. Medical examinations of each presenting registrant were carried out to determine if there were any issues that would prevent him from serving effectively in the Army. If the results of the local exam were inconclusive, the man was referred to Medical Advisory Board of specialists. Early on in the buildup, some 44% of those examined were disqualified at the local level, which led to changes in the system so that only those obviously disqualified were rejected at this point (Eanes, 1951). Also early on, dental defects were a major source of disqualification, leading to adjustments in the standards to bring this rate down.

As the war progressed, it became clear to some that physical abilities should play a greater role in how soldiers were

[29] Personnel Research in the Army II

assigned (Palmer, Wiley, & Keast, 1948). In particular, General Lesley McNair, Commander of Army Ground Forces, was concerned that men were being assigned to the Infantry who did not have the physical stamina to adequately perform their jobs. He noted results of an examination of 6,000 infantrymen in November 1943 which revealed that their average height was only 5 feet 7.7 inches, as compared to the Army average of 5 feet 8.4 inches. At his urging, a Physical Profile Plan was established in February 1944 which categorized men at reception stations into one of four categories: Profile A qualified for strenuous combat duty; Profile B qualified for less rigorous combat duty or service in ground support units; Profile C were suitable for base positions either in the U.S. or overseas; and Profile D could be used at the discretion of commanders. Although the plan was fully implemented, Palmer, Wiley, & Keast, (1948) suggest that its impact was limited due to the variety of factors other than physical ability that affected how soldiers were assigned.

Psychiatric screening was also performed at Selective Service stations and induction centers. Leading up to the war, conferences and committees associated with the NRC considered the prospect of such screening which led to Circular Letter 19 on screening for likely psychiatric problems (Whitehorn, 1951). The primary screen in this regard was a 3 to 6 minute interview conducted by a psychiatrist or physician. If deemed necessary, further tests could be conducted to assess the individual. Hunt and Stevenson (1946) indicate that frequent use was made in the Army of the *Cornell Selectee Index* (CSI), which was made up of a series of yes-no questions to determine the possible presence of "neuropsychiatric symptomatology." The authors state that, depending on the cut score used, tests such as the CSI were found to detect 50 to 90 percent of those

who were psychologically unfit for service, with only 3 to 25 percent false positives. Whitehorn (1951) indicates that some 15,000,000 men were screened by psychiatrists over the course of the war, with 1,846,000 rejected for neuropsychiatric disorders. However, many of the rejectees were later called back to serve as the need for manpower became more acute.

Individuals who passed the initial screening at induction stations were then sent to one of 36 reception centers. Here they were outfitted, given inoculations, and provided initial instruction in basic Army ways and Articles of War. It was also at reception stations that the AGCT was taken and interviews were conducted, with the information collected entered and keypunched on each soldier's Qualification Card. A classification officer was responsible for making the initial decision as to where a Soldier would be assigned, and an assignment officer saw to it that he was sent to the appropriate training center.

During 8 weeks of basic training at one of 36 Replacement Training Centers, soldiers were further tested and observed (Bingham, 1942). It is at this point that they may have been given one of the special aptitude tests described earlier, and further interviews were carried out with a personnel consultant officer. When an occupational assignment was made, a decision necessarily influenced by the needs of the Army at the time, this information was also entered on his Qualification Card. This would determine what specialty training, if any, the Soldier would be detailed to next. As described earlier, it was at this stage that candidates for OCS might also be identified and tested further. Generally, both specialty and officer training were limited to those who scored in AGCT Category I or II, with men in Categories IV and V becoming "basic soldiers."

Unit assignments were made by the Division Classification Officer based on the information contained on the Qualification Cards. These were reviewed by unit officers, company commanders, division classification officers, and personnel consultants before being finalized. Over the course of the war there were variations in the process described based on the types of men being inducted, the needs of the Army, and various other factors that demanded flexibility. And, as seen below, there were flaws in the system that rendered the seemingly logical process by which men were sorted and classified problematic, and fairly drastic shifts in personnel were required as a result.

"Illiterates"

As mentioned previously, a particular problem faced by those attempting to adequately man the Army for the war effort was the large number of those reporting for service who were functionally illiterate (a term that at that time was more acceptable than it is today), and therefore would have great difficulty absorbing training and following written directives. According to Ballantyne (2002), of the first 2,000,000 men examined for induction, 50 percent were rejected, 5 to 10 percent because of illiteracy. Clearly, the potential loss of 50,000 to100,000 men who, aside from their lack of reading skills, could be potentially effective servicemembers was substantial. As a result, research and development efforts were undertaken to (a) more accurately pinpoint men whose literacy level was so lacking that it would impair their ability to serve, and (b) fine-tune selection instruments to further identify those among this number who could be given training that would make them of use in the war effort.

Early on in the war effort, research was conducted to determine the minimum reading level required for effective service. A *Minimum Literacy Test* (MLT) was constructed in 1941, with critical scores based on the *Metropolitan Advanced Reading Test*. The MLT was then given to a sample of engineer trainees at Fort Belvoir, VA in August of that year, and scores were examined in relation to performance ratings on 14 dimensions. Researchers discovered a sharp increase in unsatisfactory ratings among men whose MLT scores put them below the fourth grade level. Therefore, this was set as the minimum literacy standard.

Initially, screening interviews were used to identify those who should be administered literacy tests. No exact guidelines were given in this regard, but it was suggested that occupational history and level of education in relation to chronological age be taken into account. When the interviewer suspected that literacy might be a problem, the MLT was given. Failure on this instrument led to the administration of the *Visual Classification Test* (VCT), a non-language group test. This measure, developed by the PRS, was made up entirely of pictures with minimal language involved in its administration. When only a few "illiterates" were being tested, or when psychologists suspected that the VCT might not be providing accurate results, a battery of other instruments was available, including a test of ability to follow instructions and a block counting test.

Near the end of 1942, the *Army Information Sheet* was implemented to identify those for whom additional literacy testing would be advisable. It was typically administered to men who had not completed the 7th grade, as well as other cases at the discretion of the interviewers or examiners. This short instrument consisted of 12 items that included writing ones

name, address, and age, as well as five paragraph comprehension questions. Anyone with less than nine items correct was administered one or more of the literacy tests cited above. However, because school achievement standards varied widely by region, problem arose with this process. So, in June 1943 a *Mental Qualification Test* (MQT) was introduced which was given to all registrants who were not high school graduates. As before, failure to pass the 17-item instrument resulted in further assessment using the standard literacy tests.

Work on literacy test refinement continued and, after trying out some 35 measures, a new test battery was introduced in June of 1944. The MQT continued to be used; however the VCT was replaced with the *Group Target Test* (GTT). This included three parts which assessed memory for motion patterns, direction sense, and spatial orientation. Registrants scoring above the criterion on the GTT were accepted to service because their performance indicated they had sufficient ability to absorb military training. Yet another screen was introduced for those who did not pass the GTT. This was an individual examination in which the registrant moved crayons up pathways of increasing complexity as the examiner counted a specific cadence. A second portion of the examination required men to complete a variety of tasks such as making patterns using blocks. For non-English speakers who failed the GTT, the follow-up examination involved looking at a series of pictures and drawing lines between items that were similar. Instructions were acted out by the examiners. Research demonstrated that this final battery of literacy tests brought about a significant increase in reliability and validity over those in use previously as indicated by performance in training.

Another non-language instrument was developed in response to requests from Special Training Units and Replacement Training Centers for a test that could be used to back up recommendations for or against discharges based on inaptitude. The test was also to be used in hospitals as an addition to other clinical techniques in making psychiatric diagnoses. Various characteristics were deemed desirable for the assessment, including that it cover the same range of abilities as the AGCT against which it would be validated, that it include both verbal and nonverbal items, that it be applicable to both black and white soldiers and be appropriate for soldiers in general, and that it involve a minimal amount of time and equipment to administer. Researchers started out with a set of 17 tests, including both commercially available instruments and those developed by the PRS staff. These were validated using 465 trainees, including approximately equal numbers of black and white soldiers. In the end, six subtests were included in AIT-1:

- Story Memory, in which the person being tested was read a paragraph which he must be able to repeat and answer questions about.
- Similarities-Differences, in which pairs of words were presented and the examinee was required to indicate how the things were similar or different. One point was given for a correct answer of each type.
- In the Digit Span subtest, a series of numbers was presented and the test taker was asked to repeat the sequence either forward or backward.
- In the Shoulder Patches subtest a pattern is presented and the person being tested must duplicate it using cut outs.

- Trail making required examinees to draw lines between numbers or letters. They were scored on how quickly the task was completed.
- Cube assembly involved presenting a picture of cubes arranged in a given pattern which must be duplicated using actual cubes. Again, time to completion determined subtest score.

The AIT was standardized using 1,000 soldiers who represented the full spectrum of AGCT scores. The tests were weighted so approximately equal contributions would be made by the verbal and nonverbal tasks. Although it is unclear how extensively the AIT was used during the war, the trail making test was adapted by researchers and practitioners in such fields as neuropsychiatry and gerontology to assess the presence and severity of various neurological disorders.

Men who would have been considered illiterate but who were judged to have sufficient capacity to absorb and apply military training were sent to Special Training Units. There they were given classroom instruction for three hours a day, with the remainder of their time devoted to military training. Reading programs were developed to accommodate various levels of starting ability, and involved materials specifically created to incorporate elements of military life. As described by Sticht (2001), the *Army Reader* included four parts. In Part 1, the men were introduced to Private Pete as he experienced various aspects of entering the Army. Part 2 addressed writing letters, while Part 3 covered issues regarding pay and allowances. Part 4 discussed the need for Army service and the goals for which the country was fighting. Other materials were created to assist in literacy training, including a monthly magazine entitled *Our War*, which included simply written descriptions of the

activities of the armed forces on various fronts throughout the world.

In the July 1943 issue of *Our War*, a letter from Private Porfirio C. Guiterrez was published in which he said, "This is my first letter in English. I have learned to read and write so that I can help protect our country." Clearly the activities undertaken to improve the skills of men so that they could effectively take part in the war effort had an impact on many levels—from imparting basic skills that would allow "illiterates" to function on a higher plane in society, to moving ASTP soldiers beyond where they ever thought they would go in pursuit of higher education.

Problems with the Classification and Assignment Process

In business, government, and the military, personnel planning can often be at odds with the pressing needs of the day. This is particularly true for the military in a time of war. The huge mobilization that took place to meet the manpower needs during World War II resulted in some anomalies that the best plans and practices could not prevent, and some which greater foresight may have helped avoid. Harrell (1945) and Palmer, Wiley, & Keast, (1948) describe some of the major issues which the Army confronted.

- For at least some part of the war, regulations dictated that a certain percentage of men assigned to the Army Air Forces (AAF) had to score 100 or above on the AGCT. This, in combination with the fact that the majority of men preferred the AAF because of higher pay and greater glamour, meant that a disproportionate number of higher scoring soldiers were assigned to this branch. According to Harrell (1945), "The AAF, even after filling all of its technical jobs, has an embarrassing abundance of men superior in GCT who are doing

routine duties as drivers, laborers, airplane handlers, etc. In a large sample 10% of AAF men in the lowest skilled job, laborer, scored in grades I and II GCT."

- As the war proceeded, the problem of distribution inequity was compounded as the need for Infantrymen increased. Between November 1944 and April 1945, 90% of the 420,000 men who were newly assigned went to the ground forces. However, between August of 1944 and April of 1945, the percentage of inductees between the ages of 18 and 25 declined from 85% to 65%. As a result, the ground forces were also forced to accept men who were less physically qualified for the arduous tasks they confronted. To counteract this problem, efforts were undertaken to improve the image of the Infantry and attract more volunteers. When these met with limited success, mandatory transfers were enacted. In March of 1944, 30,000 aviation cadets were transferred to Infantry, and later that year the process was undertaken to transfer approximately 100,000 men from the air and service forces to the ground forces. Conversely, during this same period the ground forces transferred some 22,500 men who were not physically qualified for duties in the Infantry to other jobs in the Army.
- Harrell (1945) also found fault with the occupational quota system that was put into place, which often trumped logic when it came to making assignments. He cites the example of a highly qualified clerk arriving for duty at a time when the need for mechanics was great. According to Harrell, the likelihood would be that this man would be assigned to mechanics training even

when, as it turns out, a need for clerks was found shortly thereafter. He also claims that if there was a demand for soldiers in a certain job and they weren't available, "shotgun reclassification" would take place to meet the need, meaning reassignments were made without taking into account a Soldier's background or training.

- Another problem cited by Harrell was with the rotation system. He claims that ineffective soldiers were frequently rewarded by being sent home, while those who were performing well on the various fronts around the world were being retained for combat and other duty. "In sum, contrary to the written policy, soldiers are punished by banishment from the United States when they succeed and are rewarded by return when they fail" (Harrell, 1945, p.457).
- Harrell also found fault with the officer promotion system, which he claimed was troubled by the influence of friendships between those under consideration for promotion and those making the promotion decisions. He also suggested that the requirement that unsatisfactory ratings be supported by evidence dissuaded many from even considering them.

Harrell cited other problems he saw with the personnel system, including a failure on the part of psychologists to become sufficiently knowledgeable about Army jobs and processes, and a failure on the part of Army officials to learn about the possibilities and limitations of psychology as it relates to Army functions.

Epilogue

The Allied Forces reached Berlin on April 21, 1945. On the 30th of that month, Adolph Hitler committed suicide. Two days earlier, Mussolini had been assassinated in the small village of Giulino di Mezzegra, Italy. On May 6, Germany surrendered unconditionally. However, the war in the Pacific continued until the Japanese surrendered on August 14 after the dropping of atomic bombs on Hiroshima and Nagasaki. The largest war in the history of mankind had come to an end. Some 50 million people had lost their lives, including nearly 15 million military personnel from all sides.

The planning for post-war demobilization began long before hostilities ceased. A wide variety of considerations had to be taken into account, including the need for troops in Germany, Japan, and elsewhere to ensure an orderly transition to peace; the possibility of a threat from the now-communist Russian state; and the complicated logistics entailed in bringing millions of men home from locations around the world. The initial plans called for servicemembers to be released on an individual basis taking into account several factors including length of service, whether that service involved being in combat, awards received, time spent overseas, and parenthood. Such an orderly—and relatively slow—process did not sit well with the American public, and therefore those who represent them in Congress. So, the pace of release was increased and, by the end of 1945, approximately half of the eight million soldiers had returned to civilian life. Early in 1946, the Army slowed the rate of return based on concerns that overseas obligations would be threatened. This, too, met with loud protests, including from servicemembers themselves. In response, the remaining forces were reduced by half again in the first six months of 1946. This

process continued until by June 30, 1947 the Army stood as a volunteer body of 684,000 ground troops and 306,000 airmen.

From an Army of some 270,000 just prior to the start of World War II, a force of over 8 million emerged to meet the demands of the conflict and eventually prevail. Over the course of its tenure over 15 million Americans were in uniform at one time or another—70 percent of all American men served. The colossal undertaking of ensuring that those who entered service had the capacity to be effective, and then determining where they could make the greatest contribution, did not always work exactly as planned. However, some 12 million individuals took the AGCT between 1941 and 1946, and a large proportion of them were also administered additional aptitude tests to identify those who would be effective as clerks, mechanics, radio operators, and in a variety of other roles. Research supporting these efforts was voluminous, with 40 studies alone done to determine how to best select truck drivers. Sisson (1948) summarizes much of this work. The reference list for his article includes 346 citations, all but ten of which report on work done by the Personnel Research Section of the Adjutant General's office.

For many of those who survived the war, the experience was life-changing. Whether it was taking part in battle, traveling to the far reaches of the world, or meeting men from all walks and strata of life, they were left with indelible memories and impressions. And for thousands of these men—and women—the result was more tangible. Many who entered service unable to take part in so much of what society had to offer because they couldn't read, had been given a leg up. And others who saw high school as the acme of their personal education now had

higher aspirations, that were further supported by programs such as the GI Bill.

References

Altus, W. D. (1949). Adjustment and food aversions among Army illiterates. *Journal of Consulting Psychology, 13,* 429-432.

Ballantyne, P. F. (2002). Psychology, society, and ability testing (1859-2002). Transformative alternatives to mental Darwinism and Interactionism (Chapter 5). Retrieved from http://www.igs.net/~pballan/preface(2002).html

Bingham, W. V. (1941). Psychological services in the United States Army. *Journal of Consulting Psychology, 5,* 221-224.

Bingham, W. V. (1942). The Army personnel classification system. *Annals of the American Academy of Political and Social Science, 220,* 18-28.

Britt, S. H. (1942). The Office of Psychological Personnel—report for the first six months. *Psychological Bulletin, 39,* 773-793.

Classification and Replacement Branch, The Adjutant General's Office (1945). The Army General Classification Test. *Psychological Bulletin,* 42, 760-768.

Dallenbach, K. M. (1941). The work of the emergency committee in psychology. *Journal of Consulting Psychology, 5,* 210-216.

Davis, R. A. (1943). Testing in the Army and Navy. *Journal of Educational Psychology, 34,* 440-446.

Driskell, J. E., & Olmstead, B. (1989). Psychology and the military: Research applications and trends. *American Psychologist, 44,* 43-54.

Duncan, A. J. (1947). Some comments on the Army General Classification Test. *Journal of Applied Psychology, 31,* 143-149.

Eanes, R. H. (1951). Standards used by selective service and a follow-up on neuropsychiatric rejectees in World War II. In L. Carmichael & L. C. Mead (Eds.), *Selection of military personnel* (pp. 149-156). Washington, DC: National Research Council.

Harrell, T. W., & Churchill, R. D. (1941). The classification of military personnel. *Psychological Bulletin, 38,* 331-353.

Harrell, T. W. (1945). Applications of psychology in the American Army. *Psychological Bulletin, 42,* 453-460.

Harrell, T. W. (1992). Some history of the Army General Classification Test. *Journal of Applied Psychology, 77,* 875-878.

Hunt, W. A., & Stevenson, I. (1946). Psychological testing in military clinical psychology: II Personality testing. *Psychological Review, 53,* 107-115.

Marquis, D. G. (1944). The Office of Psychological Personnel—Report for the year 1943. *Psychological Bulletin, 41,* 246-252.

Marquis, D. G. (1945). The Office of Psychological Personnel—Report for the year 1944. *Psychological Bulletin, 42,* 225-230.

Marquis, D. G. (1946). The Office of Psychological Personnel—Report for the year 1945. *American Psychologist, 42*, 346-348.

Miles, W. R. (1941). Coordination of psychological services in the national emergency. *Journal of Consulting Psychology, 5*, 216-220.

Palmer, R. R., Wiley, B. I., & Keast, W. R. (1948). *The procurement and training of ground combat troops.* Washington, DC: U.S. Government Printing Office.

Partington, J. E., & Bryant, T. R. (1946). The personnel consultant and psychological testing at Armed Forces induction stations. *American Psychologist, 1,* 110-112.

Perkins, M. E. (1955). Preventive psychiatry during World War II. In J. B. Coates, E. C. Hoff, & P. M. Hoff (Eds.), *Preventive Medicine in World War II* (pp. 172-232). Washington, DC: Office of the Surgeon General, Department of the Army.

Sisson, E. D. (1948). The personnel research program of the Adjutant General's Office of the United States Army. *Review of Educational Research, 18,* 575-614.

Staff, Personnel Research Section, Classification and Enlisted Replacement Branch, The Adjutant General's Office (1943). Personnel research in the Army: I. Background and organization. *Psychological Bulletin, 40*, 129-135.

Staff, Personnel Research Section, Classification and Enlisted Replacement Branch, The Adjutant General's Office (1943). Personnel research in the Army: II. The classification system and the place of testing. *Psychological Bulletin, 40*, 205-211.

Staff, Personnel Research Section, Classification and Enlisted Replacement Branch, The Adjutant General's Office (1943). Personnel research in the Army: III. Some factors affecting research in the Army. *Psychological Bulletin, 40*, 271-278.

Staff, Personnel Research Section, Classification and Enlisted Replacement Branch, The Adjutant General's Office (1943). Personnel research in the Army: IV. The selection of radiotelegraph operators. *Psychological Bulletin, 40*, 357-371.

Staff, Personnel Research Section, Classification and Enlisted Replacement Branch, The Adjutant General's Office (1943). Personnel research in the Army: V. The Army Specialized Training Program. *Psychological Bulletin, 40*, 429-435.

Staff, Personnel Research Section, Classification and Enlisted Replacement Branch, The Adjutant General's Office (1943). Personnel research in the Army: VI. The selection of truck drivers. *Psychological Bulletin, 40,* 499-508.

Sticht, T. G. (2001). *Learning to read with Private Pete & Sailor Sam in World War II.* Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

CHAPTER 6

1946-1973—A STABLE SITUATION IN ARMY SELECTION & CLASSIFICATION

Peter F. Ramsberger
Human Resources Research Organization (HumRRO)

Following the conclusion of World War II, inductions ended as the post-war demobilization began. However, the size of the Army was never again to dip to pre-war levels. In February 1946, the Army Deputy Chief of Staff for Personnel ordered the development of a new selection test for individual or group administration that would (a) involve a minimum amount of instruction for delivery, (b) take 30 minutes or less to administer, and (c) be as nearly equivalent as possible to the Army General Classification Test (AGCT 3a) in the lower end of the test score distribution. As described by Uhlaner and Bolanovich (1952), the Personnel Research Section (PRS) selected an existing test, known as the R-1, for this purpose. This screen was developed early in the war for administration to men with physical limitations that prevented their participation in combat or related roles, but who could potentially serve the Army in other capacities. The R-1 was a short test, composed of items from the AGCT-1a, including 17 vocabulary, 19 arithmetic, and 15 block counting questions. It was calibrated and standardized against AGCT-1a so that it would yield raw scores equivalent to standard scores on its predecessor. During the war, a standard score of 90 or above was necessary for the enlistment of men with limited physical abilities.

Beginning in April 1946, the R-1 was introduced as the screening test to be used to determine the cognitive ability of new Army recruits. The test was administered at Central Examining Stations, and a minimum standard score of 70 was required for enlistment. At this time, local recruiting stations only checked to make sure that the men who applied met requirements on such factors as age and lack of physical limitation. Those who passed this initial screen were transported to Central Examining Stations where the R-1 was administered. Based on this assessment, applicants were either accepted into the Army or rejected on the basis of the test results. Recognizing that this wasted valuable resources, the PRS set about to develop additional tests that could be used for both pre- and final screening. As described by Uhlaner and Bolanovich (1952), various factors had an impact over time in regard to the testing program development, including standards for admission into the Army, the amount of training being provided to new soldiers, whether new recruits were volunteers or draftees, and eventually, the need for a single test to be used across Services.

Development of an alternate form of the R-1 test, logically called R-2, had been undertaken prior to the end of the war, but was dropped when hostilities ceased. It comprised 17 vocabulary and 18 arithmetic items from AGCT-1b, and thus was basically a short form of that test. The R-2 was standardized in 1942 using the AGCT-1c, and again in 1946 using the AGCT-3a. At that time the decision was made to study the R-2 further while developing new forms R-3 and R-4. These latter tests were made up of 10 reading and vocabulary, 10 arithmetic reasoning, and 15 pattern analysis items which were drawn from questions formulated for, but not used in, AGCT 3a through 3d. All three new forms were administered experimentally in April and May of 1946 to four groups. All completed the AGCT-3a, with one each taking the R-2, R-3, R-

4, and both R-3 and R-4. The findings suggested that the R-3 and R-4 were superior to the R-2 in regard to their correlation with the AGCT-3a, and that a 25 minute time limit yielded better results than either 15 or 20 minutes. Noting that the reliabilities were adequate, the R-3 and R-4 were put into operational use in August 1946 at the Central Examining Stations, with the R-2 used as a prescreen at local recruiting offices.

Early in 1948, responsibility for final screening of volunteers shifted from Central Examining Stations to Main Recruiting Stations. At that time it was decided that the R-3 and R-4 should be replaced, so AGCT-1c and 1-d were republished as R-5 and R-6. When new editions were issued. they were named the General Classification Tests (GCT) 5 and 6. With the new tests in place, the R-3 and R-4 were used for prescreening at the local level. This system remained in place until January 1, 1950 when the Armed Forces Qualification Test (AFQT) became operational.

Armed Forces Qualification Test

The Selective Service Act of 1948 specified that, "The passing requirement for the General Classification Test shall be fixed at seventy points." At the time the bill passed, the GCT was in use by the Army and the Air Force for screening purposes. The Navy was using the Navy Applicant Qualification Test, and the Marine Corps forms 1c and 1d of the AGCT,. In addition, three different methods were used for converting and reporting scores on the tests. With the new law specifying a cut score of 70 on the GCT, it became apparent within DoD that a single test for qualifying applicants across Services would be desirable. A committee made up of representatives from each of the Services was established on January 27, 1949 to start planning for this eventuality. The Army was given prime responsibility for the project through the

Personnel Research Section, Personnel Research and Procedures Branch of the Adjutant General's Office. Their charge was to develop a global measure of cognitive ability using vocabulary, arithmetic reasoning, and spatial relations items. The test, to be called the Armed Forces Qualification Test (AFQT), was to have maximum sensitivity in the Army Standard Score range of 60 to 90 (the range in which cut scores were likely to be set), but adequately cover the entire spectrum of ability. Because the test would potentially be used to screen both volunteers and inductees, it needed to be validated in reference to the full mobilization population rather than the population of those serving in peacetime under non-draft conditions. A uniform scoring system was to be used, and relationships to previous screening tests determined so that the qualitative makeup of the Services over time could be measured and documented. Specific test items were to be related to everyday military activities so as to be accessible to those taking the test. Verbal instructions were to be kept as simple as possible, and the importance of speed in performance should be minimized so as not to penalize test takers who may have relatively high cognitive ability but work at a slower pace.

The vocabulary portion of the test included logical associations (e.g., He is *ill*. Test taker chooses between *hurt*, *pale*, *sick*, *sad*). The arithmetic reasoning items were created so as to minimize the verbal element in responding. There were seven types of items, including fractions, ratios, percentages, and weights and measures. Spatial questions addressed item identification, folding and unfolding of patterns and forms, and identifying which whole figure could be constructed from parts given. In all, 554 items were created and divided into two forms. These were administered to 7,000 Army, Navy, and Air Force recruits along with the operational screening tests in use at the time. Cases with missing data were eliminated, and the sample was adjusted to reflect the

racial proportions in the World War II reference population. This reduced the number of cases to 5,742.

To examine item difficulty, an index was computed by assigning "the lowest reference test score interval at which an item was answered correctly by at least 50% of the group in the interval" (Uhlaner & Bolanovich, 1952, p. 14). Nine difficulty levels resulted. Correlations between the vocabulary items on the new test and those on the reference tests were computed and were in the .45 range. Arithmetic item validities were similar, ranging from .39 to .52 depending on the reference test in use. Average item validities for the spatial relations portion of the test were somewhat lower (.26 to .35). Correlations were also run between the scores on items of one type with the total score on the items of the other two types to determine the level of independence of the three portions of the test. The results suggested that the spatial relations items were most different (correlations from .21 to .39), while there was a stronger relationship between the arithmetic reasoning and vocabulary items (.27 to .56). This was attributed to the fact that it was impossible to completely eliminate the verbal element in the arithmetic items.

Based on the results obtained from the experimental administrations of the test, researchers settled on 90 items, 30 each vocabulary, arithmetic reasoning, and spatial relations. The experimental results indicated that 85% of Service members completed this number of items in 45 minutes, which was set at the new time limit. Two criteria were used for selecting specific items for the two forms of the AFQT—difficulty level and consistency in difficulty level across experimental samples. On each portion of the test items were selected so that 15 were in the lower half of the difficulty spectrum, 5 in the middle level, and 10 in the higher end of the distribution. An "over-all" index was then

computed for selecting the specific questions to be included on the two forms of the test. This was done by computing, "a weighted composite of item biserial correlation coefficients equal to the sum of the three validity coefficients plus the sum of the two internal consistency coefficients minus the sum of the two independence coefficients" (Uhlaner and Bolanovich, 1952, p. 18).30 In addition to this index, researchers took care to select items that had the highest validity and internal consistency and maximum independence. They also examined the percent passing index and performed a careful review to make sure that the two forms of the AFQT were similar in content and psychological process.

In standardizing AFQT scores, the decision was made to use the population of all men on active duty as of December 1944. The assumption was that no major population changes had occurred in the intervening period and that this tactic would be more economical than testing a sample of the entire current civilian population. After converting AFQT scores from the experimental sample to Army Standard Score values, the distributions were "blown up" to the reference population of 11,694,229 Service members on duty in 1944. "A composite cumulative percentile curve was set up in 5-point Army Standard Score intervals, and the percentage of the total distribution was calculated for each interval" (Uhlaner and Bolanovich, 1952, p. 20).

Four samples of current servicemembers were then selected, including all new recruits at three Army, one Air Force, and two Navy installations in July 1949. Two of the samples took AFQT-1 and two AFQT-2. Both took the reference test (AGCT-1c, AGCT-

[30] The three validity coefficients correspond to the three types of items on the test, the two internal consistency coefficients represent the two forms of the test, and the two independence coefficients reflect the comparisons of each set of item types with the other two items types.

1d, Army Classification Battery). Within AFQT groups, one sample took the AFQT first and the reference test second, with the order reversed in the other sample. After scoring the tests (score equals number of correct answers), statistics were computed to determine if test order had an impact on outcomes, and whether there were notable differences between the two forms. The answer was no in both cases. Using equipercentile conversion, AFQT scores were translated into Army Standard Scores. When the distributions were compared with what was expected, researchers found that there was a close match at the top end of the scale (Standard Score above 80), but there were higher percentages in the lower ranges than anticipated. The explanation for this outcome was that achievement on the test was influenced by both cognitive ability and literacy skills. This led to the recommendation that non-verbal tests be used in conjunction with AFQT to get a better assessment of the true abilities of those lacking reading proficiency. Further statistical analyses revealed that AFQT was highly correlated with (a) AGCT (.90 or greater); (b) Army Classification Battery (ACB) Aptitude Area I tests of reading, vocabulary, arithmetic, and pattern analysis (.92); and (c) years of education (.69).

On January 1, 1950, AFQT-1 and AFQT-2 were put into operational use at Navy and Joint Army-Air Force recruiting stations. The R-3 and R-4 were used at local recruiting stations for pre-screening so that those unlikely to meet the minimum cut score on the AFQT (percentile 13, raw score 31) would not be sent forward for further processing. Once in operational use, an anomaly was found in the results being achieved. In Army recruiting stations, those being tested were given the AFQT and the ACB. . It was found that a large percentage of those passing the AFQT were scoring below the equivalent Standard Score on the Aptitude Area I ACB tests. This occurred despite the fact that

the ACB was developed using the old AGCT-3a and 3b tests and was found in experimental usage to correlate .92 with AFQT. Researchers labeled this phenomenon "operational slippage," and attributed it to a lack of standardization in the way in which the AFQT was administered at the recruiting stations. To empirically determine that this was the case, the Assistant Chief of Staff for the G-1 ordered a study be carried out.

A sample of 1,000 men was drawn from the nearly 5,000 who were processed at four Army reception stations during the week of February 12, 1961. The sample was selected to duplicate the proportions of men in 10-point standard score groupings among the World War II population. These men were administered the alternate form of the AFQT from the one taken during screening, this time taking it under standardized conditions in their training divisions. The same score standardization procedures described above were followed using these data. What researchers found was that the raw score equivalents to Army Standard Scores were similar in the original and follow-up studies where the test was administered under standard conditions. However, the results for Standard Score 80 and below were markedly different under operational conditions. For instance, where a raw score of 31 translated to a Standard Score of 70 in both instances when the test was given under standard conditions, a raw score of 39 was equivalent to a Standard Score of 70 when the test was administered operationally. This led researchers to conclude that "Standard Score conversions for AFQT, then, did not hold up at the levels below Standard Score 80 for AFQT scores obtained under operational administration conditions" (Uhlaner & Bolanovich, 1952, p. 26).

To address these concerns, the responsibility for AFQT administration was transferred to commissioned personnel

psychologists assigned to the Armed Forces Examining Stations. A two-week course was developed to train these individuals in proper test administration procedures, and Uhlaner and Bolanovich (1952) indicate that at the time they issued their report, "preliminary evidence points in the direction of considerable improvement" as indicated by a reduction in the variance between results seen in controlled and operational administrations of the AFQT.

The AFQT was used by all Services for aptitude screening until 1972. New forms were introduced periodically, as follows:

Year Introduced	Forms
January 1, 1950	1 & 2
January 1, 1953	3 & 4
August 1, 1956	5 & 6
July 1, 1960	7 & 8

Items were written for AFQT Forms 9 and 10, but with the introduction of the Armed Services Vocational Aptitude Battery (ASVAB), they were not needed.

Some changes were introduced to the test over time. For instance, starting with Forms 3 and 4, a Tool Knowledge test was added. The purpose was to reduce the literacy requirements of the test and decrease the correlation with years of education. Maier (1993) describes the process of making the transition from one AFQT form to the next:

> Whenever a new form of the AFQT was introduced, the score distributions were carefully examined to make sure that no abrupt changes in

> the proportions of examinees in the various score categories occurred. No problems with the calibration were found, and the tests continued in use until they were replaced. (p. 69)

It should be noted that women seeking entry into the Army were originally tested using the AFQT. This situation changed in 1953 with the introduction of the Armed Forces Women's Selection Test (AFWST). This was done because research demonstrated that the spatial and tool knowledge items of the AFQT had an adverse impact on women. That is, it was less likely women would qualify for service because of these items, even though it was unclear that they were related to their subsequent training or likely job duties. Therefore, the AFWST measured similar domains as the AFQT, namely verbal skills, arithmetic reasoning, and pattern analysis, but it was constructed to place greater emphasis on assessing those skills related to clerical and administrative positions which were the principal domain of military women during this period. The test was used until 1976, with some modifications over time, at which point all potential military personnel were tested with the ASVAB. An AFQT composite is computed by combining scores from relevant ASVAB subtests which is equated to previous versions of the AFQT, thereby enabling DoD to track scores across time and Services.

Army Classification Battery

As described in Chapter 5, during World War II the Personnel Research Section of the Adjutant General's office undertook a wide variety of studies to develop tests that would enable the classification of men into occupational specialties for which they had the necessary aptitude to succeed in training and on the job. In 1949, ten of these were selected to be put into

operational use as the Army Classification Battery (ACB). They were:

1. Reading and Vocabulary (RV)
2. Arithmetic Reasoning (AR)
3. Pattern Analysis
4. Mechanical Aptitude (MA)
5. Army Clerical Speed (ACS)
6. Army Radio Code Aptitude (ARC)
7. Shop Mechanics (SM)
8. Automotive Information (AI)
9. Electric al Information (EI)
10. Radio Information (RI)

Along with the tests, a set of ten composites was developed for prediction of performance in ten groups of related occupations known as Aptitude Areas (AAs). It was acknowledged that the composites were not based on an abundance of validity data, and that validation work would be ongoing.

Over the course of the next decade, researchers undertook a slew of such validation studies focusing on which combinations of ACB tests were most valid for predicting success in specific jobs or job families. (e.g., Hodges, Brogden, & Uhlaner, 1955; Ransone, Brogden, & Uhlaner, 1956, Hodges, Brogden, & Uhlaner, 1956). With the knowledge gained from these endeavors and the realization that Army jobs were changing over time, researchers began looking at reconstituting the AAs (Zeidner, Harper, & Karcher, 1956). Existing data used for this purpose came from 42 studies of 62 groups of enlisted personnel either in training or on the job. Researchers examined the 10 unweighted 2-test composites with the highest validity for a given job family. Their goal in identifying job families was to make them as homogenous as possible, and keep the

number low. Their goal in regard to composites was to pick that combination of tests with the highest psychometric and face validity, while reducing the intercorrelations between the composites themselves. One approach taken to reduce the intercorrelations was to double weight the test in a given composite that was most directly related to a job family (e.g., the general maintenance composite was 2 Shop Mechanics + Pattern Analysis). The results of this work led to the reduction of the Aptitude Areas from ten to seven. In addition, the ACB composites were all combinations of just two ACB tests, as opposed to previous composites which included from two four.

This research also suggested a change in the tests themselves. Results indicated that Electrical Information (EI) and Radio Information (RI) were lacking enough easy items to be able to discriminate between ability levels at the lower end of the aptitude spectrum. To address this problem, a new Army Electronics Information Test was developed to replace EI and RI (Trump, Klieger, White, & Karcher, 1957). It included picture items that required the examinee to associate pictured objects and demonstrate knowledge of how they function electronically. Verbal items required demonstration of familiarity of electronic principles. The goal was to be able to identify a level of expertise that was obtained due to interest in the subject area, as opposed to formal electronics training.

Another subject that received a good deal of attention throughout the 1950s was the potential for developing measures that would predict combat effectiveness. Several early studies, two of which were conducted in Korea, sought to isolate those characteristics evident in soldiers who were identified by supervisors and peers as effective in this environment. (Tieman, Campbell, Goldstein, Johnson, & Yaukey, 1952; Egbert, et al.,

1957). Subsequent research investigated use of content from personality measures examined in the Korea studies that appeared to have the most potential, along with a "test" of masculine and outdoor interests, and a variety of other experimental tests. In the end it was determined that the best predictors were the personality measure (named the Classification Inventory) and that assessing masculine interests, the General Information Test. New composites were then formed for infantry occupations (IN) that used a double weighted Classification Inventory combined with the Arithmetic Reasoning test from the ACB. Another Aptitude Area was created for artillery, armor, and combat engineer jobs (AE). Its composite combined the General Information Test and the existing ACB Automotive Information Test (Willemin & Karcher, 1958). The resulting Aptitude Areas and Composites remained in effect for the next decade.

A longitudinal research program was undertaken in the 1960s to assess whether the composites in use were still viable given the changing nature of Army jobs (Maier & Fuchs, 1969). The study included some 25,000 men in over 100 different training courses. The existing Aptitude Areas were retained, and different composites of ACB scores were used to determine that combination which best predicted success in training as indicated by course scores. A key consideration was that the new composites not be overly complicated so that they would be simple enough to use in the field. Tests within each composite were weighted to reflect their importance in predicting success in training in that Aptitude Area. Each composite contained a measure of general learning ability to ensure that no one job family was receiving a disproportionate share of men low on this dimension.

Another research program was undertaken in the mid 1960s, this time evaluating the Aptitude Areas and examining whether existing ACB tests should be modified and/or new tests added to provide greater predictive power in regard to training success (Maier & Fuchs, 1972a; Maier & Fuchs, 1972b). This resulted in a significant overhaul of the Aptitude Areas, with Infantry, Field Artillery, Electronics, Clerical and General Maintenance retained. Motor Maintenance was broadened somewhat to become Mechanical Maintenance; General Technical was replaced with Skilled Technical; Radio Code was dropped; and Operators and Food, and Surveillance and Communications were added. In addition, a variety of changes were introduced to the ACB tests, with some dropped (e.g., Shop Mechanics, Clerical Speed), others shorted/updated (e.g., Word Knowledge, Automotive Information), and yet others added (e.g., Trade Information, Mathematics Knowledge). A new set of composites was also developed containing from two to five ACB tests per Aptitude Area.

As described in Chapter 7, the mid-1970s saw the move to the ASVAB as the joint-service selection and classification test. This was implemented in January of 1976. Despite the discontinued use of the ACB at that time, it seems clear that the work of researchers over some two decades made a monumental contribution to the science of occupational classification that not only had a strong impact on the development of the ASVAB, but also on the field as a whole.

References

Egbert, R. L., Meeland, T., Cline, V. B., Forgy, E. W., Spickler, M. W., & Brown, C. (1957). *FIGHTER I: An analysis of combat fighters and non-fighters* (Tech. Rep. No. 44). Washington, DC: Human Resources Research Office.

Hodges, C. B., Brogden, H. E., & Uhlaner, J. E. (1955). *Prediction of success in the field wireman course* (PRB Tech. Research Note No. 46). Washington, DC: Personnel Research Branch, Adjutant General's Office.

Hodges, C. B., Brogden, H. E., & Uhlaner, J. E. (1956). *Prediction of success in personnel and administration courses* (PRB Tech. Research Note No. 56). Washington, DC: Personnel Research Branch, Adjutant General's Office.

Maier, M. H. (1993). *Military aptitude testing: The past fifty years* (Tech. Rep No. 93-007). Monterey, CA: Defense Manpower Data Center, Personnel Testing Division.

Maier, M. H., & Fuchs, E. F. (1969). *Development of improved Aptitude Area Composites for enlisted classification* (Tech. Research Rep. No. 1159). Arlington, VA: U.S. Army Behavioral Science Research Laboratory.

Maier, M. H., & Fuchs, E. F. (1972a). *An improved differential Army classification system* (Tech. Research Rep. No. 1177). Arlington, VA: U.S. Army Behavior and Systems Research Laboratory.

Maier, M. H., & Fuchs, E. F. (1972b). *Development and evaluation of a new ACB and Aptitude Area system* (Tech. Research Note No. 239). Arlington, VA: U.S. Army Behavior and Systems Research Laboratory.

Ransone, A. L., C. B., Brogden, H. E., & Uhlaner, J. E. (1955). *Prediction of success in armament maintenance courses* (PRB Tech. Research Note No. 66). Washington, DC: Personnel Research Branch, Adjutant General's Office.

Tiemann, R. S., Campbell, J. T., Goldstein, L. G., Johnson, C. D., & Yaukey, D. W. (1952). *Validation of personnel measures against combat performance of enlisted men in Korea. IV. Experimental aptitude tests* (Tech. Research Rep. No. 948). Washington, DC: Personnel Research Branch, Adjutant General's Office.

Trump, J. B., Klieger, W. A., White, R. K., & Karcher, E. K. (1957). *The Army Electronics Information Test* (Tech. Research Rep. No. 1096). Washington, DC: Personnel Research Branch, Adjutant General's Office.

Uhlaner, J. E., & Bolanovich, D. J. (1952). *Development of Armed Forces Qualification Test and Predecessor Army Screening Tests, 1946-1950* (PRS Rep. No. 976). Washington, DC: Personnel Research Branch, Adjutant General's Office.

Willemin, L. P., and Karcher, E. K. (1958). *Development of combat Aptitude Areas* (PRB Tech. Research Rep. No. 1110). Washington, DC: Personnel Research Branch, Adjutant General's Office.

Zeidner, J., Harper, B., & Karcher, E. K. (1956). *Reconstitution of the Aptitude Areas* (PRB Tech. Research Rep. No. 1095). Washington, DC: Personnel Research Branch, Adjutant General's Office.

CHAPTER 7

MOVING TO THE ALL-VOLUNTEER FORCE (1973-1982)

Wayne S. Sellman
Human Resources Research Organization (HumRRO)

Move to the All-Volunteer Force

Throughout most of American history, the military has been composed of volunteers. However, conscription was the primary means of obtaining military personnel during World Wars I and II and the Korean Conflict to the point that its renewal became perfunctory. The decision to move to an all-volunteer military evolved from criticism of the inequities of conscription during the Vietnam War – who shall serve when not all serve? In the late 1960s, President Richard Nixon established a commission to develop a comprehensive plan for eliminating conscription and move toward an all-volunteer force. The commission built a case for a volunteer military by pointing out the unfairness of conscription, establishing the feasibility of a volunteer force on economic grounds, and by refuting all major arguments against ending conscription and relying totally on volunteers (President's Commission on an All-Volunteer Armed Force, 1970).

The commission (known as the Gates Commission for its chairman, former Secretary of Defense Thomas S. Gates) believed that sufficient numbers of qualified youth could be persuaded to volunteer by increasing military pay to levels to be more competitive with civilian wages. They disputed claims

that total reliance on volunteers would lead to a mercenary force consisting mainly of racial/ethnic minorities, the poor, and the uneducated, and that the country would be losing the "social contract" between its young people and the state (i.e., that youth from all walks of life would miss out on the opportunity to serve). After much debate within the Nixon Administration, Congress and across the country, it was decided that an all-volunteer force was feasible, affordable, and would not jeopardize the nation's security. Thus, the authority for conscription was allowed to lapse on July 1, 1973. The last draft call went out in December 1972. On June 30, 1973, Dwight Elliot Stone, a 24-year old apprentice plumber from Sacramento, California, because the last person to be inducted into the armed forces. The Defense Manpower Commission (1976) and Lee and Parker (1977) provide comprehensive documentation regarding establishment of the all-volunteer force and analysis of the factors which led to elimination of the draft.

Time has proved the viability of the voluntary force concept, but experience has taught that there are fundamental Defense policies and programs needed to guarantee its success.

- Ensure military pay is relatively competitive with civilian pay.
- Establish and maintain a highly-motivated, professional recruiting force with a marketing orientation.
- Initiate recruiting advertising to communicate the benefits and opportunities of service, and to reinforce a positive public image of the military.
- Establish enlistment incentives that feature job skill training, choice of assignment locations, money for college, and bonuses for longer terms of service and entry into hard-to-fill occupations.
- Investigate changes in military personnel utilization (e.g., use of civilians, women, and lower-aptitude personnel) to reduce demand for scarce high-quality men.
- Improve retention of careerists through quality-of-life initiatives with a focus on maintaining a family orientation.

In the early and mid-1970s, the president and Congress provided sufficient funds to make the all-voluntary military work. Then, in the late 1970s, military pay fell behind civilian pay, there was little money for recruitment advertising, and education benefits and enlistment bonuses were cut sharply. As a result, the Military Services encountered great difficulty attracting high-quality young people to consider the military as a post-high school option. In recruiting, you get what you pay for. As the per-capita funding (dollars per recruit) rises, so does the quality of new enlistees; and when the investment falls, so does recruit quality. Recognizing that relationship, in the early 1980s, the Reagan Administration and Congress restored the necessary levels of recruiting resources.

The American people historically have opposed conscription and large standing armies. The use of conscription to raise our forces has been the exception rather than the rule. With adequate resources and support to attract and retain the brightest personnel, conscription is not needed to meet future military personnel requirements. An all-volunteer force is more expensive than a conscription force in terms of military compensation and funds for advertising and enlistment incentives. However, a voluntary military is less expensive in

overall costs. It is more stable and career-oriented, thereby leading to extra performance and experience, with reduced training and other turnover costs. During conscription, 10 percent of new inductees reenlisted; today's new recruits reenlist at a rate of approximately 50 percent. In short, military service is an economically rational choice for high-quality men and women looking for an edge on life. The military also is a good choice for people who want to serve a greater cause.

For readers interested in more detail on the transition from conscription to a volunteer military, we recommend a 2006 book by former Under Secretary of Defense for Personnel and Readiness Bernard Rostker entitled *I Want You: The Evolution of the All-Volunteer Force.* This book presents a history of the development of today's volunteer military along with a review of the major policy issues and research undertaken to support Department of Defense decision makers over the past 30 years.

Impact of the All-Volunteer Force on Selection and Classification

As noted in Chapter 6, in 1950, the Services returned to a single test, the Armed Forces Qualification Test (AFQT), to be used in conjunction with the reinstitution of the Selective Service System draft. The new AFQT was modeled after the Army General Classification Test (AGCT) and basically measured the same variables (i.e., general learning ability). However, unlike the AGCT and the aptitude tests of World War I, the AFQT was specifically designed to be used as a screening device (Eitelberg, Laurence, Waters, & Perelman, 1984). Thus, the AFQT was established for the purpose of both (a) measuring examinees' general ability to absorb military training within a reasonable length of time so as to eliminate those who did not possess such ability, and (b) providing a uniform measure of

examinees' potential usefulness in the service, if qualified on the test (Uhlaner & Bolanovich, 1952).

Whereas the AFQT was used to identify individuals who had a reasonable probability of success in service, other Service-specific tests were required for job classification. The Army Classification Battery, the Navy Basic Test Battery, and the Airman Qualification Examination, just to name a few, were used from the late 1950s to the mid 1970s (Waters, 1997). During this period, the AFQT was administered to military applicants (including draft inductees) at Armed Forces Examining and Entrance Stations (AFEES) across the country for selection purposes.[31] If individuals successfully "passed" the AFQT and qualified on the other enlistment standards (e.g., physical fitness, moral character), they were accepted for service and sent to basic training, although the specific occupation to which they would be assigned had not yet been determined. During basic training, new enlistees were administered their Service's classification test and were assigned to appropriate military occupations.

Origin of the Armed Services Vocational Aptitude Battery

In 1972, during the run-up to the all-volunteer force, Roger T. Kelley, Assistant Secretary of Defense for Manpower and Reserve Affairs, issued a policy that allowed each Service to discontinue use of the AFQT and to replace it with Service-specific selection tests. [32] During field visits, recruiters would frequently report to him that many qualified applicants were lost

[31] Since women were not subject to the draft, a different aptitude test was used for female applicants for enlistment. The Armed Forces Women's Selection Test was administered to female applicants in lieu of the AFQT from 1956 to 1974.

[32] In spite of staff objections, Kelley also questioned the need for the Joint-Service AFEES, which administered the AFQT for entry into service, but he stopped short of abolishing them.

to service because they "flunked" the AFQT (Lee & Parker, 1977). Kelley's line of reasoning is given in extracts from the memorandum he sent to the Services in May 1972 (Kelley, 1972).

> Each Service is to provide this office with the means intended for use in expressing test results, if other than the AFQT, in terms of mental categories. Upon providing satisfactory evidence of your ability to do this, you will be relieved of the requirements to follow test procedures prescribed by this office.
>
> It continues to be a function of the Office of the Secretary of Defense (OSD) to monitor Service practices to be sure that the quality mix of new entries is appropriate and that you are neither screening out qualified candidates nor admitting unqualified ones. It will be the responsibility of each Service to employ a combination of tools, including but not limited to written tests, which achieve that result most effectively.
>
> For the record, this action does not reflect an anti-test bias. It does reflect the conclusion that you know, or ought to know more about how to select people for your respective Service than we in OSD do. Given the opportunity to employ your own selection tools, you will not be surprised to be held strictly accountable for your results in recruiting.

This policy was implemented over the objections of Kelley's staff, who believed that there should be some Defense-wide measure of recruit quality (Lee & Parker, 1977).

From 1973–1976, each Service was permitted to develop conversion tables from its own selection test as a basis for estimating an individual's AFQT scores. (AFQT scores were used as a means of reporting recruit quality to Congress across time and across Services. That function continues today.) In May 1974, William K. Brehm, who replaced Kelley as Assistant Secretary of Defense for Manpower and Reserve Affairs, reversed Kelley's Service-specific selection testing policy. In a memorandum to the Services, Brehm (1974) directed that a single test battery be used by the Services for selecting enlistees and for placing them into military occupations. Brehm cited five advantages to a common test as the basis for his decision (Sellman, 1980).

- High school testing would be more useful to all Services.
- Applicants for more than one Service would not be subjected to multiple testing.
- Inter-Service referrals of applicants would be improved.
- More accurate comparisons of cross-Service applicants would be enhanced.
- Test development work of Service psychologists would be concentrated on a single enlisted accessions and classification instrument.

By combining selection and classification testing at the AFEES, the testing process was to be made more expedient for the newly implemented all-volunteer military. Young people applying for enlistment would take one test and come away

from the AFEES knowing not only if they qualified for enlistment, but, if qualified, also the military occupation to which they would be assigned. Thus, the new policy enabled the Services to improve the matching of applicants with available occupations and allowed job guarantees for those individuals qualified for enlistment.

The Armed Services Vocational Aptitude Battery (ASVAB) was selected for this purpose since (a) it was already being used in the Department of Defense High School Testing Program and (b) at the time, the Air Force and Marine Corps were administering a parallel form of ASVAB in their own operational testing programs. The version of ASVAB used for high school testing was ASVAB-2. The Air Force and Marine Corps used ASVAB-3, a parallel form to the high school version, for operational testing of applicants. The Armed Forces Qualification Test (AFQT), comprising verbal, mathematical, and spatial perception measures, would be derived from the ASVAB. The AFQT would again be the primary enlistment screen for all Services. The Air Force, already serving as the executive agent for the High School Testing Program, was designated as the executive agent for further ASVAB development and expansion.

Establishment of the ASVAB Steering Committee and Working Group

In the same May 1974 memorandum that directed the use of ASVAB as the Joint-Service selection and classification test, Brehm established the ASVAB Steering Committee. Chaired by Donald W. Srull, Deputy Assistant Secretary of Defense for Manpower Requirements and Analysis, the Steering Committee was composed of senior officers and civilians from the offices of their respective Deputy Chiefs of Staff for Personnel. The

main function of the Steering Committee was to provide policy recommendations on ASVAB development, implementation, and use to the Assistant Secretary of Defense for Manpower and Reserve Affairs. The Steering Committee also conveyed the positions of their Services on ASVAB issues to the Office of the Secretary of Defense (OSD).

In turn, the Steering Committee chartered the ASVAB Working Group – which continues today as the Joint-Service Manpower Accession Policy Working Group. The responsibilities of this body were to design the new ASVAB so as to accommodate Service requirements and to develop plans for validating the test. (Working Group responsibilities later evolved into resolving ongoing problems in ASVAB research and development, implementation, and maintenance.) While the Working Group was to be a Joint-Service activity, its members were to represent the positions of, and be responsible to their individual Services.

Mr. Gus C. Lee, OSD Special Assistant for the All-Volunteer Force, was appointed by the Steering Committee as chairman of the Working Group. The group was composed of Service testing policy staff officers and scientists from the Service personnel research laboratories. Because the Air Force was executive agent for ASVAB, the Air Force Human Resources Laboratory was designated as the lead research activity. Three forms of ASVAB were to be developed. ASVAB-5 would be used in the DoD High School Testing Program while ASVAB-6/7 would be administered as the common selection and classification battery. Assistant Secretary Brehm's original timetable for ASVAB implementation was September 1, 1975.

Early Problems with ASVAB Development and Implementation

Most of the information contained in the following sections about ASVAB development and implementation is taken from *History of the Armed Services Vocational Aptitude Battery: 1974-1980*. This report was prepared by the ASVAB Working Group in 1980, because of the widespread interest in the test after Robert B. Pirie (Assistant Secretary of Defense for Manpower, Reserve Affairs, and Logistics) notified the House and Senate Committees on Armed Services about the ASVAB misnorming and the erroneous enlistment of large numbers of low-ability recruits. The section of that report on the origin of ASVAB was written by then Major W. S. Sellman, based on DoD and Service letters, memoranda, reports, and research documents, as well as personal experience. In 1980, at the time of the report, Sellman was assigned to the Office of the Secretary of Defense where he had access to all the relevant documents. Previously, Sellman served as the Air Force testing policy staff officer (HQ, USAF) and chairman of the ASVAB Working Group.

The first major issue to arise within the ASVAB Working Group was introduced by the Navy and involved the implementation of ASVAB-6/7 before they had been validated against success in Service training schools. The Air Force and Marine Corps had already been using ASVAB-3 as their selection and classification test, so validation before implementation was not a real concern for those Services. The Army agreed with the Navy that validation was important but was willing to accept statistical correspondence between existing and new ASVAB tests as evidence of the test's validity.

The Navy representatives on the Working Group felt so strongly about this issue that they elevated it to the Steering Committee. In a November 6, 1974 memorandum, Rear Admiral E. J. Carroll, Assistant Chief for Personnel Planning and Programming, Bureau of Naval Personnel, wrote to his Service counterparts on the Steering Committee. "Recent meetings and discussions concerning progress in developing ASVAB-5/6/7 have cast considerable doubt on prospects for full, effective implementation of the new batteries by September 1, 1975. . . . . There appears to be no reasonable possibility that adequate Service validation can be accomplished to permit exclusive use of ASVAB for all Service selection and entry processing purposes as of September 1, 1975." Admiral Carroll enclosed a projected "plan of action and milestones" based on what he believed to be realistic estimates of time still required to resolve significant ASVAB development problems, plus the time required for validation by the Services. He recommended that implementation be delayed until June 1, 1976 (Carroll, 1974).

Admiral Carroll's memorandum did not generate strong support for the Navy's position among the other Services. Major General George W. Putnam, Army's Director of Military Personnel Management, responded that while the Navy's proposed changes did not present any problems to the Army, he believed that the proposal should be given to the Working Group for its consideration before presenting it for a decision by the Steering Committee (Putnam, 1974). Brigadier General Kenneth McLennan, Director for Marine Corps Manpower Plans and Policy, wrote, "The Marine Corps favors the September 1975 implementation of ASVAB Forms 5, 6, and 7 for accession testing . . . The rationale for this position stems from the fact that since July 1, 1974, the Marine Corps has been using ASVAB Form 3 as its principal instrument for accession

testing and is acutely desirous of obtaining backup tests for ASVAB-3 as early as possible" (McLennan, 1974).

The Air Force, as executive agent for ASVAB, could hardly go against the OSD implementation guidance. Major General Kenneth L. Tallman, Director of Personnel Plans, Headquarters Air Force, replied to Admiral Carroll stating, "We fully understand your reservations concerning the implementation of ASVAB without appropriate Navy validation. One solution to your problem might be to explain your misgivings to OSD. In this regard, a suggestion that you be allowed to continue administration of your basic classification battery along with ASVAB-5 until you have collected sufficient data to complete validation research might be appropriate. In any event, because of our previous experience with ASVAB and the OSD pressure for its early adoption as a common production test, we feel compelled to adhere to the plan for September 1975 implementation (Tallman, 1974).

Despite the Service positions on the Navy's proposed new implementation date, the issue went to the ASVAB Steering Committee for consideration. Because of slippages in test item development, the Air Force Human Resources Laboratory had already fallen approximately 45 days behind schedule. After discussion of what this slippage would mean for the September 1975 implementation, Admiral Carroll brought up the Navy position that, because of the lack of validation data, implementation should be delayed until June 1976. Mr. Srull then directed Mr. Lee to prepare a report to Assistant Secretary Brehm which would discuss problems in meeting the September 1975 implementation, as well as suggest alternative courses of action and provide Steering Committee recommendations (Srull, 1975).

The report was submitted to Assistant Secretary Brehm in February 1975. In his transmittal memorandum, Srull indicated his impatience with the Services for not getting ASVAB ready in a more expeditious manner. It should be noted that, almost from the time of the May 1974 decision to use ASVAB as a Joint-Service selection and classification test, both Brehm and Srull had frequently indicated to OSD staff and representatives of the Air Force that they could not understand why it took so long to develop a test. As a personal anecdote, the Air Force testing policy staff officer, Major W. S. Sellman, who later replaced Gus Lee (upon his retirement) as chairman of the ASVAB Working Group, was called on to brief Srull on the status of ASVAB development. At the end of the briefing, Srull banged his hand on his desk in great frustration and said, "It takes longer to build a test than to put a man on the moon." Ultimately, Assistant Secretary Brehm approved a delay in ASVAB implementation to October 1, 1975.

More Service Concerns – More Delays

For about the next six months test development was basically "on-track." Then, in July 1975, with camera ready masters of ASVAB-5/6/7 already at the printers, the Army member of the Steering Committee (Major General John F. Forrest, Director of Military Personnel Management) indicated that the test was too difficult and recommended that easier items be substituted for the more difficult ones in order to permit selection among applicants at lower levels of ability. The Steering Committee agreed and requested the Army to provide substitute items for inclusion in the test. Once this decision was made, the Air Force stopped the printing.

With such an "eleventh-hour" situation concerning the printing, it would have been impossible to meet the milestones necessary

for an October 1975 start-date. Accordingly, on July 31, 1975, in a memorandum to the Service assistant secretaries for manpower and reserve affairs, Assistant Secretary Brehm changed the ASVAB implementation date to January 1, 1976 (Brehm, 1975). The new date would coincide with the date that all DoD enlistment testing would be done at Armed Forces Examining and Entrance Stations. Although it did not agree with the January 1976 implementation date, the Navy decided to comply with Brehm's direction (Watkins, 1975).

The item difficulty issue was not easy to resolve. Upon receipt of the substitute items from the Army, the Air Force "cut" them into the camera ready master copies and again initiated printing. In August 1975, the Navy lodged a formal protest with OSD because they believed that the Steering Committee and OSD had made a unilateral decision to include the Army's easy items at the expense of the Navy. The Navy's position was that the easy items would not differentiate among their personnel in the upper ability range and would therefore adversely impact its classification system.

Ultimately, test items acceptable to all Services were selected by the Working Group for inclusion in ASVAB. The Working Group solution was formally endorsed by the Steering Committee on August 21, 1975. As it turned out, this Steering Committee meeting was the last one that would be held for three years. Mr. Srull had left his position within OSD, so the August 1975 meeting was chaired by Mr. I. M. Greenberg (Acting Deputy Assistant Secretary of Defense for Manpower Requirements and Analysis) and Mr. Gus C. Lee, chairman of the Working Group, had retired from OSD in May 1975. During the August 1975 meeting, Major W. S. Sellman, Air Force testing policy staff officer, was asked by the Steering

Committee to assume that position. Major Sellman served as the chairman of the Working Group from August 1975 to August 1978.

The last policy decisions made by the Steering Committee in 1975 concerned (a) the difficulty of items to be included in the operational ASVAB, and (b) issues related to the high school version of the test. Additionally, the Air Force, as executive agent, was directed to proceed as soon as possible to develop the follow-on versions – ASVAB-8/9/10. The policy matters now having been resolved, OSD believed that the efforts remaining were technical and that there was no further need for its involvement in the process.

Implementation of ASVAB as the Joint-Service Enlistment Test

From August through December 1975, no significant ASVAB issues arose. The time was filled with frenetic activity as the Services and their personnel research laboratories were faced with and solved last-minute technical and logistical problems. After the early inter-Service disagreements, this period was marked by Service cooperation as they worked together to effect the implementation of ASVAB. That there were problems because of the compressed time schedule in which many technical development tasks occurred will be evident in a later section of this chapter. Be that as it may, ASVAB was implemented as the DoD enlistment eligibility test on January 1, 1976.

AFQT Categories

For convenience in reporting, AFQT scores are expressed on a percentile scale and grouped into five broad categories. Those scoring in Categories I and II tend to be above average in

cognitive ability; those in Category III, average; those in Category IV, below average; and those in Category V markedly below average. (Category III is sometimes divided at the 50th percentile into subcategories A and B. This facilitates reporting the proportion of scores above and below the mean of the AFQT distribution.) By law, Category V applicants and those in Category IV who have not graduated from high school are not eligible for enlistment.

Table 7.1 shows the percentile scores for the various categories and the percent of the military personnel in World War II who fell into each category. From the introduction of AFQT in the early 1950s until new norms were implemented in 1982, the DoD enlistment tests, including ASVAB, were normed against the so-called World War II mobilization population. (This World War II norming population included the 11,694,229 men under arms as of December 31, 1944). One consequence of the ASVAB misnorming, which will be discussed in the next section, was development of new norms based on a nationally representative sample of contemporary youth to replace the 30-year old norms. Today, ASVAB is normed against a nationally representative sample of young people ages 18 to 23 years who tested in 1997 as part of the Bureau of Labor Statistics' National Longitudinal Survey of Youth (Segall, 2004).

Table 7.1. AFQT Scores by Category World War II Reference Population

AFQT Category	Percentile Score Range	Percent of WWII Reference Population in each Category (Officer and Enlisted)
I	93 – 100	8
II	65 – 92	28
III	31 – 64	34
IV	10 – 30	21
V	1 – 9	9
	TOTAL	**100%**

The ASVAB Misnorming

The ASVAB was introduced for DoD-wide use in January 1976. During its development, the Armed Forces Qualification Test (AFQT) part of ASVAB was normed against the AFQT portion of Service selection and classification tests then in use. Norming is simply a method through which a test's raw scores are converted to percentile scores. Raw scores on a test are by themselves of very limited usefulness. They must be normed against the scores of a defined and relevant population. This enables standardized scores from different versions of a test (percentiles in the case of the AFQT) to have the same interpretive meaning. For example, a percentile score of 65 from the current version of AFQT should equate to a percentile score of 65 from the AFQT used during the 1960s or 1980s. For the AFQT, the norms allow the Department of Defense to evaluate its new recruits across time and across Services. If the norms inaccurately translate raw scores into percentile scores,

test users cannot accurately compare the capabilities of new enlistees with those who served in the past.

Shortly after the implementation of ASVAB, the Services found that, compared to previous experience, an excessive number of people were scoring in the upper two AFQT categories (percentiles 65 – 99) as determined by the AFQT portion of ASVAB. An adjustment was made in July 1976 to the test calibration that reduced the numbers in those upper categories. There was no evidence at that time which indicated a change in norming was needed for the bottom half of the score range. Consequently, the norming for the lower half of the AFQT score distribution was left essentially intact.

In a report published in April 1978 by the Center for Naval Analyses (CNA), data were provided for Marine Corps recruits which indicated that the changes made during 1976 to reduce the numbers who scored in AFQT Categories I and II were not accurate – that, in fact, more high-aptitude recruits were entering service than was being reported (Sims, 1978). The CNA report also said that more low-aptitude people were actually being enlisted than reported. Service psychologists reviewed findings from this report and other Service analyses, and concluded that no changes should be made in scoring the test.

Because of the potential impact on military personnel, the Office of the Secretary of Defense (OSD) asked CNA to repeat its study. Preliminary results from the second study, issued in May 1979, differed from the findings of the first. CNA now suggested that upper ability levels were being accurately measured, but that far more low-scoring recruits were being enlisted than had been shown in its previous study (Sims & Truss, 1979).

Confronted with these conflicting findings, in May 1979, OSD asked the Army Research Institute for the Behavioral and Social Sciences (ARI) to conduct a new study using applicants (not simply Marine Corps recruits) for all Services. Further, because of the uncertainty of the results of the two CNA analyses and the need for accuracy in the AFQT scores, it was considered prudent to obtain another estimate of the AFQT norms, one conducted outside of the Department of Defense and with a sample that had no significant experience with, or incentive to pass DoD tests. Thus, the Educational Testing Service (ETS) was asked to perform a companion study using high school juniors and seniors. The results of these two additional studies were expected to provide DoD with a sounder basis for evaluating test norming errors. The ARI (Maier & Grafton, 1980) and ETS (Boldt, 1980) analyses were completed in June 1980. In addition, CNA reanalyzed its data (Sims & Truss, 1980). The three studies agreed that many individuals who had been classified within AFQT Category III should have been placed in Category IV instead.

Because of the importance of the test norming issue, OSD retained the services of three nationally known testing experts to review the final results of the ARI, ETS, and CNA analyses. Dr. Robert L. Linn, University of Illinois; Dr. Richard M. Jaeger, University of North Carolina; and Dr. Melvin R. Novick, University of Iowa, completed their assignment in June 1980 and advised OSD that the norms in use were inaccurate and had inflated the AFQT scores of individuals in the lower end of the test score range. In addition, they believed that of the three analyses, the ARI report, using Service applicants, had produced the most appropriate results. Consequently, the consultants recommended that the current norms be replaced by the ARI norms (Jaeger, Linn, & Novick, 1980).

Application of the new ARI norms had a substantial impact on the score distributions of Service recruits. Table 7.2 shows the percentage of enlisted accessions during FY 1979 that scored in the various AFQT categories, both as originally reported using the operational norms and using the ARI norms. The significant changes were the decrease of the percentages in Category III and the increase in Category IV. In the Army, the reported percentage of personnel in Category IV changed from 9 percent to 46 percent. As a result of the ASVAB misnorming, the proportion of new recruits scoring about the 50th percentile on the AFQT dropped from over 60% in 1975 to just over 30% in 1976, with the percent scoring at or above the 50th percentile not returning to the pre-1976 level until the mid-1980s.

Table 7.2. AFQT Scores By Military Service, FY 1979 Reported and Corrected Scores (Percent) *

	AFQT Categories				
	I	II	III	IV	TOTAL
DoD					
Reported	4	25	66	5	100
Corrected	3	25	42	30	100
Army					
Reported	3	17	70	9	100
Corrected	2	15	37	46	100
Navy					
Reported	6	29	62	4	100
Corrected	5	33	44	18	100
Marine Corps					
Reported	3	22	72	4	100
Corrected	2	23	49	26	100
Air Force					
Reported	6	35	60	0	100
Corrected	5	36	50	9	100

* Numbers may not add due to rounding.

Profile of American Youth and the Implementation of ASVAB Forms 8, 9, and 10

To ensure currency, the DoD periodically introduces new enlistment tests. Because the old versions of ASVAB had been in use for almost four years, the new forms were a natural progression of test development. In fact, the Office of the Secretary of Defense (OSD) had requested the Air Force, as ASVAB executive agent, begin development of the new forms before the ASVAB misnorming had been discovered. Thus, there were replacement forms of the test available for implementation when the corrected norms were ready to be put into place. On October 1, 1980, the Department of Defense introduced three new forms of the ASVAB (Forms 8, 9, and 10).

As a result of the misnorming episode, OSD, in conjunction with the Services, decided to develop contemporary norms for ASVAB-8/9/10. The ASVAB was administered to a nationally representative sample of approximately 12,000 youth ages 18 to 23 years participating in the U.S. Department of Labor's National Longitudinal Survey of Youth. Known as the Profile of American Youth (PAY), this study marked the first time a vocational aptitude test had been given to a national probability sample of young people. Undertaken to calculate updated norms for the ASVAB, the PAY sample replaced the World War II norming population (Sellman & Valentine, 1981; U.S. Department of Defense, 1982) and provided a better understanding of the quality and representativeness of new enlistees compared to the general youth population (Sellman & Doering, 1982; Sellman & Laurence, 1981).

Women, Minorities, and the All Volunteer Force

The military has historically been a predominantly white, male institution. However, the need for additional "manpower" in wartime repeatedly brought minorities and women into the Services (Holm, 1982; Manning & Griffith, 1998; Segal, 1989), often in separate units with restrictions on the types of positions that could be held. During World War I, Native Americans served in integrated units (Bernstein, 1991). Following World War II, the Services began integrating blacks into the ranks upon orders by President Harry Truman (MacGregor, 1981). Almost 30 years later, limits on the percentage of women permitted in the Services were lifted. Soon after, the military entered the all-volunteer force era.

Women in the Military

There was a two-percent ceiling on the proportion of women allowed in the military between 1948 and 1967. During this period, women made up approximately one percent of the force. Following the elimination of the ceiling in 1967, little change occurred in the number of women seeking entrance into the Armed Forces. Through the early 1970s, the number of female Service members gradually increased but remained small (U.S. Department of Defense, 1993).

Two significant events profoundly affected the growth rate of women in the military: the expanding role of women in society (spurred by the women's movement) and the transition to the all-volunteer force. At the start of the volunteer force in 1973, there were four primary reasons for increasing the number of women in the military:

1. This action would compensate for a projected decline in the 18-year-old male market.

2. The pool of young women was a vastly underutilized resource.
3. The Services could significantly increase the quality of the force since each additional high-quality female recruit meant one less lower-quality male. In the early 1970s the concern was that the Services might be unable to recruit the desired level of high-quality males.
4. Changing social norms and economic conditions were resulting in greater proportions of women going into the workforce, with many entering what were described as nontraditional female positions.

The increase in women in the military since 1972 brought about significant changes across all aspects of personnel management: in training programs and physical fitness regimens, in assignments, in living arrangements, and in medical services. It also created new administrative issues regarding pregnancy, the proportion of single parents in the military, child care arrangements during peacetime and deployment, and joint spouse marriages (where husband and wife both serve in uniform).

The most controversial issue, however, was defining the role of women in combat (Landers, 1989). The conflict concerning the role of women in the military is a complex issue. The Presidential Commission on the Assignment of Women in the Armed Forces studied the matter in-depth and provided its report to President George H.W. Bush on November 15, 1992 (*The Presidential Commission*, 1992). The position taken by Commission members in the final report and reaction to that report indicate that the issue of women in combat continued to be a source of debate.

On April 28, 1993, then Secretary of Defense Les Aspin issued a Service-wide policy on the assignment of women in the

Armed Forces which directed the Services to open more specialties and assignments to women. Specifically, more aircraft, including aircraft engaged in combat missions, were to be opened to qualified women. The Navy was instructed to open as many ships to women as possible and to submit a proposal to repeal the combat exclusion law to allow women on ships engaged in combat missions. Army and Marine Corps officials were requested to consider additional opportunities for women in field artillery and air defense artillery. Additionally, Aspin's policy (1993) established an implementation committee "to ensure that the policy on the assignment of women is applied consistently across the Services, including the Reserve Components" (p. 2). The committee was charged with reviewing parental and family policies, pregnancy and deployability policies, and the "Risk Rule" which was issued in 1988 by Secretary of Defense Frank Carlucci. This rule stated that

> Risks of direct combat, exposure to hostile fire, or capture are proper criteria for closing noncombat positions or units to women, when the type, degree, and duration of such risks are equal to or greater than the combat units with which they are normally associated within a given theater of operations. If the risk of noncombat units or positions is less than comparable to land, air, or sea combat units with which they are associated, then they should be open to women.

Review of the Risk Rule led to policy changes concerning women in combat as described in the Direct Ground Combat Definition and Assignment Rule (Aspin, 1994; Perry, 1994).

The implementation of the Direct Ground Combat Rule in the 1990s—opened more specialties and positions to females than before—providing for broader career paths for Servicewomen. This, in turn, improved chances for promotion to the highest ranks.

Since the advent of the All Volunteer Force in 1973, the percentage of female accessions has nearly tripled, rising from 5 percent to approximately 16 percent in 2005 (U.S. Department of Defense, 2007). While the Services have increased their proportions of women, the proportions and numbers are not comparable to female representation in the civilian population. Reasons for the difference include lower inclination of women than men to apply for and enter the military, combat exclusion constraints, and Service policies. Department of Defense Youth Polls indicate that young women are approximately one-half less inclined to join the military than young men (U.S. Department of Defense, 2007; Ramsberger, 1993).

Although much progress has been achieved with regard to gender equity, much work remains. The representation of women has increased and many previously closed positions have been opened to women. The military continues to consider current and future roles for women in uniform (U.S. Department of Defense, 2006).

Minorities in the Military

During deliberations on the feasibility of ending the draft, several questions regarding fairness emerged. One such question involved racial representation – would a volunteer force draw excessively upon black recruits? The Gates Commission projected that the racial composition of the force would not be fundamentally changed by ending the draft (President's Commission on an All-Volunteer Armed Force, 1970). It considerably underestimated

participation by blacks. The percentage of Black enlisted accessions increased, with some fluctuation, following the end of the draft. Dramatic increases in the proportion of Black accessions coincided with the misnorming of the ASVAB which led to erroneous enlistment of many low-scoring applicants. Thus, representation of blacks – whose test scores are generally lower than those of whites – increased during the misnorming period. When, by the mid-1970s, the proportion of black recruits doubled the Gates Commission projections, supporters of conscription claimed to have found solid evidence for over-representation of minorities in the Services because they perceived few viable education or civilian employment opportunities. This issue contributed to a growing discussion in the late 1970s and early 1980s of returning to the draft as a way of fostering more equitable representation and fairness in sharing the burden of military service.

In the early 1980s, revised standards corrected the ASVAB scoring error. Following the implementation of the new ASVAB norms, the proportion of black recruits returned to approximately the same level as before the test scoring error. A gradual increase in black accessions ensued through the early 1990s; participation has remained relatively stable into the 21st century (U.S. Department of Defense, 2007).

Hispanics make up a much smaller, but growing, proportion of the Armed Forces than blacks. Hispanics in the military comprised about five or six percent of non-prior service accessions in the early 1970s. A slight decrease occurred in Hispanic recruits in the early 1980s after the ASVAB norms were corrected. By the late 1980s, the Services had returned to pre-volunteer force levels of Hispanic accessions. There has been a steady increase in enlisting men and women of Hispanic descent ever since (U.S. Department

of Defense, 2004). In 2005, approximately 14 percent of new enlisted recruits were Hispanic (U.S. Department of Defense, 2007).

While the increase in the number of Hispanic servicemembers since the advent of the volunteer military mirrors the increase of this ethnic group in the civilian population, it continues to lag in representation. The Services keep pace with the growth but have not been able to close the gap. The primary reason why Hispanics are underrepresented in the Armed Forces may be that youth from Hispanic homes have much higher high school dropout rates (22 percent) than students of other racial/ethnic groups (ranging from 3 to 10 percent; Laird, DeBell, Kienzi, & Chapman, 2007). Relatively few non-high school graduates are accepted by the military. To enlist, high school dropouts must have either an alternate credential (e.g., GED or college course credit) or higher enlistment test scores than applicants with diplomas. Thus, there are fewer opportunities for individuals that fail to complete traditional high school programs.

Recruits of other races (i.e., American Indian or Alaskan Native, Asian, Native Hawaiian or Pacific Islander, or two or more races) accounted for less than 10 percent of all non-prior service enlisted accessions in 2005 (U.S. Department of Defense, 2007). Compared to their representation at the end of the draft era (less than one percent), however, non-black, non-Hispanic minority youth have increased their representation within the military nearly ten-fold.

In the Armed Forces, there are similar proportions of other minorities as there are in the civilian population. However, comparisons differ within race. American Indians, Alaskan Natives, Native Hawaiians, and Pacific Islanders enlist in greater proportions than are found in the civilian population. But, Asian-

Americans and people of two or more races are less likely to join one of the Services than to pursue civilian opportunities (e.g., higher education, employment).

The racial and ethnic composition of the non-prior service enlisted accessions closely mirrors the civilian population of 18-24 year old youth. Hispanics tend to be slightly underrepresented and blacks somewhat overrepresented. Overall, the Military Services provide opportunities to individuals of all races and ethnicities. In concert with the color-blind distribution of the benefits of joining the military, those who volunteer to bear the burden of serving their country do so regardless of skin color or ethnic background.

Non-Cognitive Testing

In general, the military uses two criteria to determine the success of first-term enlistees: (a) success in basic and job skill training and subsequent performance on the job, and (b) completion of an initial obligated period of service. While ASVAB, a measure of cognitive abilities, is a valid predictor of training and on-the-job performance, it has considerably less utility in predicting first-term attrition. Attrition has grown with the all-volunteer military to the point that today approximately one-third of new recruits fail to complete a three-year term of service. While some recruits are lost because of medical disabilities and other non-pejorative causes, most are dismissed on more grievous grounds including inaptitude, behavioral disorders, motivational problems, criminal behavior, drug use, financial irresponsibility, and other signs of misconduct.

Replacing recruits who leave service early is a costly and time-consuming process. Consequently, over the past 40 years, the Services have conducted an extensive research campaign to improve the prediction of first-term attrition. Perhaps the first and most important work in this area was accomplished by Dr. Eli

Flyer (1959) at the Air Force Personnel Research Laboratory, who documented the direct relationship between high school graduation and military adjustment. Flyer concluded that "The most dramatic way to reduce unsuitability discharge would be to require a high school diploma from all Air Force recruits." Since that time, this relationship between high school graduation status and first-term attrition has been corroborated by the other Services and has stood the test of time.

Enlightened by such findings, the Services soon incorporated this research into their personnel screening processes. Beginning with the Air Force in 1961 and culminating with the Navy and Marine Corps in 1965, education criteria were used in conjunction with aptitude test scores to screen individuals for military enlistment (Laurence, 1993). Individuals without high school diplomas were required to achieve higher test scores to qualify for entrance. By combining aptitude and educational achievement, the Services hoped to select from among applicants those who were more trainable and who had better changes of successfully fulfilling their contracted terms of enlistment (Laurence 1997; Sellman, 1986).

While level of education, as an enlistment standard, has proved to be a useful means to control first-term attrition, the Services have continued their quest to improve adaptability screening through the use of non-cognitive measures. Early research focused on biodata assessment which is generally described as the collection of demographic and past biographical information (e.g., education credentials, school history, work experience, and personal characteristics) from applicants (Mael, 1994). Later research would consider measures of temperament and personality such as reliability, emotional stability, and

acceptance of authority (Edwards, McBride, Waters, & Laurence, 1993).

The Services examined biodata as a selection device to complement aptitude testing and other available criteria. Collecting such data from potential recruits would serve as an additional enlistment hurdle. That is, the ASVAB, a cognitive measure, would be used as the primary screening tool. Applicants that meet minimum enlistment test standards would then be subject to additional screening (e.g., physical, moral character, non-cognitive). A detailed review of biodata and its use in personnel selection can be found in Laurence & Waters (1993).

Each of the Armed Services began using some type of biodata instrument in recruit screening as early as the 1970s. Following the release of a General Accounting Office report in 1982, then Secretary of Defense Caspar Weinberger established a Joint-Service committee – the Adaptability Screening Group (ASG) – to explore the use of biodata measures to reduce attrition in the Services. The goal of the ASG was to develop a DoD-wide biodata tool – The Adaptability Screening Profile (ASP). To that end, the ASG examined three biodata inventories that individual Services were developing or using. These included the Air Force's History Opinion Inventory (HOI), the Navy's Recruiting Background Questionnaire (RBQ), and the Army's Military Applicant Profile (MAP) (Trent, 1993).

Some items from the RBQ and MAP were selected for the preliminary Armed Services Applicant Profile (ASAP) item pool. Additional items to tap areas not addressed were included. A series of reviews, trials, analyses, and refinements led to two 50-item alternate forms of the ASAP that were to be used in conjunction with a 70-item form of the temperament inventory,

the Assessment of Background and Life Experiences (ABLE) (Trent, 1993; White, Nord, Mael, & Young, 1993). However, just as the ASAP was ready for operational use in July 1989, it was tabled for another approach – the Compensatory Screening Model (CSM) - to serve as a selection tool to address attrition.

The thrust of the Compensatory Screening Model project was to devise alternative criteria that would expand the manpower pool by opening up enlistment eligibility to individuals with alternative education credentials (e.g., GED certificates) or non-graduates while minimizing first-term attrition. This would be accomplished by allowing positive individual attributes in an applicant's record to compensate for an education credential with a relatively high attrition risk. The list of valid attributes turned out to be relatively short: educational attainment, cognitive aptitude, age at enlistment, marital status, and number of dependents (Dempsey, Fast, & Sellman, 1977).

The only promising variable in addition to these more familiar ones was the Armed Services Applicant Profile. The Services were wary of the ASAP, however, because its susceptibility to faking, coaching, and other forms of response distortion threatened its validity in an operational environment. Consequently, DoD-wide implementation of a CSM was suspended due to lack of consensus among the Services about its long-term usefulness (McBride, 1993). The Navy was the only Service to implement a CSM as part of its enlistment screening. The model did not include the ASAP and was only used for applicants without a high school diploma. However, without the ASAP, the CSM provided minimal benefit. As a result, the CSM project was largely abandoned for the same reasons that operational use of the ASAP was cancelled.

As of the mid-1980s, none of the non-cognitive measures mentioned above had been implemented. The Services

expressed concern that biodata and temperament instruments were subject to faking by applicants and coaching by recruiters. There was also the issue of validity – such instruments proved to be valid during research, but how would they work when enlistment depended on the actual answers given? Would recruits only provide what they perceived as socially acceptable answers? In addition, the Services were reluctant to introduce a biodata or temperament screening measure because they viewed it as just another hurdle for recruits. While they recognized that such measures might well reduce first-term attrition, the Services also saw their use as having the potential to reduce the size of the manpower pool from which new recruits are drawn, thereby making recruiting more difficult.

While non-cognitive testing for military personnel selection and job classification was not introduced into the DoD enlistment testing system in the 1970-1980s, research continued. A more detailed discussion about contemporary non-cognitive research is contained in Chapter 2.

References

Aspin, L. (1993, April). Memorandum for Secretaries of the Military Departments: Policy on the Assignment of Women in the Armed Forces. Washington, DC: Secretary of Defense.

Aspin, L. (1994, January). Memorandum for Secretaries of the Military Departments: Direct Ground Combat Definition and Assignment Rule. Washington, DC: Secretary of Defense.

Bernstein, A. (1991). *American Indians and World War II: Toward a new era in Indian affairs*. Norman, OK: University of Oklahoma Press.

Brehm, W. K. (1975, July). Memorandum for Assistant Secretaries of the Military Departments (Manpower and Reserve Affairs): Enlisted Accession Processing. Washington, DC: Assistant Secretary of Defense (Manpower and Reserve Affairs).

Brehm, W. K. (1974, May). Memorandum for Assistant Secretaries of the Military Departments (Manpower and Reserve Affairs): Armed Services Vocational Aptitude Test Steering Committee. Washington, DC: Assistant Secretary of Defense (Manpower and Reserve Affairs).

Boldt, R. F. (1980). *Check scaling of the AFQT 7C portion of the Armed Services Vocational Aptitude Battery Form 7, and General Classification Test Form 1C to the Armed Forces Qualification Test scale*. Princeton, NJ: Educational Testing Service.

Carroll, E. J. (1974, November). Memorandum for ASVAB Steering Committee: ASVAB Development and Implementation. Washington, DC: Assistant Chief for Personnel Planning and Programming, Bureau of Naval Personnel.

Defense Manpower Commission (1976). *Defense manpower: The keystone of national security*. Report to the President and the Congress. Washington, DC: Author.

Dempsey, J. R., Fast, J. C., & Sellman, W. S. (1977). A method to simultaneously reduce involuntary discharges and increase the available manpower pool. Paper presented at the Office of the Secretary of Defense/Office of Naval Research Conference on First-Term Enlisted Attrition, Leesburg, VA. In H. W. Sinaiko (Ed.), *First term enlisted attrition. Vol. I: Papers*. Washington, DC: Smithsonian Institution.

Edwards, J. E., McBride, J. R., Waters, B. K., & Laurence, J. H. (1993). Adaptability screening: Conclusions and implications. In T. Trent & J. H. Laurence (Eds.). *Adaptability screening for the Armed Forces* (pp. 215-228). Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel).

Eitelberg, M. J., Laurence, J. H., Waters, B. K., & Perelman, L.S. (1984). *Screening for service: Aptitude and education criteria for military entry*. Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Installations, and Logistics).

Flyer, E. S. (1959). *Factors relating to discharge for unsuitability among 1956 airmen accessions to the Air Force* (WADC-TN-59-201). Lackland AFB, TX: Air Force Personnel Research Laboratory.

Holm, J. (1982). *Women in the military: An unfinished revolution*. Novato, CA: Presidio Press.

Jaeger, R. M., Linn, R .L., & Novick, M .R. (1980). *A review and analysis of score calibration for the Armed Services Vocational Aptitude Battery*. Washington, DC: Committee Commissioned by the Office of the Secretary of Defense.

Kelley, R. T. (1972, June). Memorandum for Assistant Secretaries of the military departments (Manpower and Reserve Affairs): Qualitative standards in moving to a volunteer force. Washington, DC: Assistant Secretary of Defense (Manpower and Reserve Affairs).

Landers, R. K. (1989). Should women be allowed into combat? *Congressional Quarterly's Editorial Research Reports*, *2*(14), 570-582.

Laird, J., DeBell, M., Kienzi, G., & Chapman, C. (2007). *Dropout rates in the United States: 2005* (NCES 2007059). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Laurence, J. H. (1997). Does the education credential still predict attrition? In J. M. Arabian (Chair), *Everything old is new again: Current research issues in accession policy*. Symposium conducted at the 105th Annual Convention of the American Psychological Association, Chicago, IL.

Laurence, J. H. (1993). Education standards and military selection: From the beginning. In T. Trent & J. H. Laurence (Eds.), *Adaptability screening for the Armed Forces* (pp. 1-40). Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel).

Laurence, J. H., & Waters, B. K. (1993). Biodata: What's it all about? In T. Trent & J. H. Laurence (Eds.), *Adaptability screening for the Armed Forces* (pp. 41-70). Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel).

Lee, G. C., & Parker, G. Y. (1977). *Ending the draft: The story of the All Volunteer Force* (FR-77-1). Alexandria, VA: Human Resources Research Organization.

MacGregor, M. (1981). *Integration of the Armed Forces: 1940-1965*. Washington, DC: Center of Military History.

Mael, F.A. (1994). If past behavior really predicts future, so should biodata's. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 273-291). Hillsdale, NJ: Erlbaum.

Maier, M. H. (1993). *Military aptitude testing: The past 50 years* (TR-93-007). Monterey, CA: Defense Manpower Data Center.

Maier, M. H. & Grafton, F. (1980). *Renorming ASVAB 6 and 7 at Armed Forces Examining and Entrance Stations.* Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Manning, L. & Griffith, J. E. (1998). *Women in the military: Where they stand* (2nd ed.). Washington, DC: Women's Research and Education Institute.

McBride, J. R. (1993). Compensatory Screening Model Development. In T. Trent & J. H. Laurence (Eds.). *Adaptability screening for the Armed Forces* (pp. 163-214). Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel).

McLennan, K. (1974, November). Memorandum for Rear Admiral E. J. Carroll: ASVAB development and implementation. Washington, DC: Director of Manpower Plans and Policy, HQ U.S. Marine Corps.

Perry, W. (1994, July). Memorandum for Secretaries of the Military Departments: Application of the definition of Direct Ground Combat and Assignment Rule. Washington, DC: Secretary of Defense.

Presidential Commission on the Assignment of Women in the Armed Forces: A Report to the President. (1992). Washington, DC.

Putnam, G. W. (1974, November). Memorandum for Rear Admiral E. J. Carroll: ASVAB development and implementation. Washington, DC: Director of Military Personnel Management, HQ Department of the Army.

Ramsberger, P. F. (1993). *Influences on the military enlistment decision-making process: Findings from the 1991 youth attitude tracking study* (FR-PRD-93-06). Alexandria, VA: Human Resources Research Organization.

Rostker, B. (2006). *I want you: The evolution of the All Volunteer Force*. Santa Monica, CA: The RAND Corporation.

Segal, D. R. (1989). *Recruiting for Uncle Sam: Citizenship and military manpower policy*. Lawrence, KS: University of Kansas Press.

Segall, D. O. (2004). *Development and evaluation of the 1997 ASVAB score scale*. Monterey, CA: Defense Manpower Data Center.

Sellman, W. S. (1986). Military adaptability screening: A manpower management perspective. In A. R. Lancaster (Chair), *Recent developments in military suitability research*. Symposium conducted at the 28th Annual Conference of the Military Testing Association, Munich, FRG.

Sellman, W. S. (1980). Origin of the ASVAB. In ASVAB Working Group (Eds.), *History of the Armed Services Vocational Aptitude Battery: 1974 – 1980*. Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics).

Sellman, W. S. & Doering, Z. D. (1982, August). The Profile of American Youth: Background and description of the 1980 nationwide administration of the Armed Services Vocational Aptitude Battery. In W. S. Sellman (Chair), *The Profile of American Youth: Results and implications*. Symposium conducted at the 90th annual convention of the American Psychological Association, Washington, DC.

Sellman, W. S. & Laurence, J. H. (1981, October). Aptitude testing in DoD and the Profile of American Youth Study. In W. S. Sellman (Chair), *Profiling the aptitudes of the current youth population*. Symposium conducted at the 23rd annual conference of the Military Testing Association, Washington, DC.

Sellman, W. S. & Valentine, L. D. (1981, August). *Aptitude testing, enlistment standards, and recruit quality*. Paper presented at the 89th annual convention of the American Psychological Association, Los Angeles, CA.

Sims, W. H., & Truss, A (1980). *A reexamination of the normalization of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 6, 7, 6E, and 7E* (CNS 1152). Alexandria, VA: Center for Naval Analyses.

Sims, W. H., & Truss, A. (1979). *A reexamination of the normalization of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 6A, 6E and 7E* (Memorandum No. 79-3059). Alexandria, VA: Center for Naval Analyses.

Sims, W. H., & Truss, A. (1978). *An analysis of the normalization and verification of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 6 and 7* (Study No. 1115). Alexandria, VA: Center for Naval Analyses.

Srull, D. W. (1975, February). Memorandum for the Assistant Secretary of Defense (Manpower and Reserve Affairs): Implementation of New ASVAB as Single Entry Test. Washington, DC: Deputy Assistant Secretary of Defense (Manpower Requirements and Analysis).

Tallman, K. L. (1974, November). Memorandum for Rear Admiral E. J. Carroll: ASVAB Development and Implementation. Washington, DC: Director of Personnel Plans, HQ U.S. Air Force.

Trent, T. (1993). The Armed Services Applicant Profile (ASAP). In T. Trent & J. H. Laurence (Eds.). *Adaptability screening for the Armed Forces* (pp. 71-100). Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel).

U. S. Department of Defense, Defense Human Resources Activity. (2007, June). *Youth propensity and influencer attitudes: Youth poll and advertising tracking findings*. Washington, DC: Joint Advertising, Market Research, and Studies.

U. S. Department of Defense, Office of the Assistant Secretary of Defense (Force Management and Personnel) (1993). *Population representation in the Military Services: Fiscal year 1992*. Washington, DC: Author.

U. S. Department of Defense, Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics) (1982). *Profile of American Youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery*. Washington, DC: Author.

U. S. Department of Defense, Office of the Under Secretary of Defense (Personnel and Readiness) (2004). *Population representation in the Military Services: Fiscal year 2002*. Washington, DC: Author.

U. S. Department of Defense, Office of the Under Secretary of Defense (Personnel and Readiness) (2006). *Population representation in the Military Services: Fiscal year 2004*. Washington, DC: Author.

U. S. Department of Defense, Office of the Under Secretary of Defense (Personnel and Readiness) (2007). *Population representation in the Military Services: Fiscal year 2005*. Washington, DC: Author.

U. S. Government Printing Office (1970). *The report of the President's commission on an all volunteer armed force*. Washington, DC: Author.

Uhlaner, J. E. & Bolanovich, D. J. (1952). *Development of the Armed Forces Qualification Test and predecessor Army screening tests, 1946 – 1950* (Report No. 976). Washington, DC: Personnel Research Section, Department of the Army.

Waters, B. K. (1997). Army Alpha to CAT-ASVAB: Four-score years of military personnel selection and classification testing. In R.F. Dillon (Ed.), *Handbook on testing*. Westport, CT: Greenwood Press.

Watkins J. D. (1975, October). Memorandum for Commander, Navy Recruiting Command: Enlisted Accession Testing. Washington, DC: Assistant Chief of Personnel Planning and Programming, Bureau of Naval Personnel.

White, L. A., Nord, R. D., Mael, F. A., & Young, M. C. (1993). In T. Trent & J. H. Laurence (Eds.). *Adaptability screening for the Armed Forces* (pp. 101-162). Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel).

CHAPTER 8

PROJECT A: ONE OF A KIND

Wayne S. Sellman and John P. Campbell
Human Resources Research Organization

This chapter describes personnel selection and classification research on a scale never attempted before in terms of (a) the types and variety of information collected, (b) the number of jobs that were considered simultaneously, (c) the size of the samples, and (d) the length of time that individuals were followed as they progressed through the organization. What follows in this chapter provides the background to the political mandate behind the Joint Service Job Performance Measurement/Enlistment Standards Project, and the Army's response: Project A.

The Origins of Project A

The events that helped shape Army personnel policy began in the 1970s, at the end of the Vietnam War. In 1973, the draft came to an end and the All Volunteer Force was instituted. By 1975, first-term attrition had reached 26.6% among high school graduate enlistees and 51.4% among non-high school graduate enlistees. Although the size of the Army had been reduced drastically from the Vietnam War era, the high attrition rates placed an enormous burden on recruiting. These times were best summarized in a now famous Department of the Army white paper on the "Hollow Army" (Meyer, 1980). The Army was also beginning the largest force modernization program since World War II which would place increased skill demands on

enlisted personnel in the context of a significant decline in the number of eligible youth—a decline that was projected to begin about 1982 and continue through 1996.

In March 1980, Richard Danzig, Principal Deputy Assistant Secretary of Defense for Manpower, Reserve Affairs, and Logistics, presented the Defense Manpower Overview Statement to the House and Senate Committees on Armed Services. In this testimony, the Committees were advised that the military enlistment test, the Armed Services Vocational Aptitude Battery (ASVAB), had been misnormed, with the result that scores in the lower ranges were artificially inflated. Approximately 360,000 young men and women, who had entered service during the period 1976 through 1980, would have been otherwise unable to meet enlistment standards (Eitelberg, 1988). As a consequence of the misnorming, members of Congress and policymakers in the Department of Defense (DoD) became interested in the methods used to set enlistment standards and to establish recruit quality requirements.

In Congress' view, the fact that the ASVAB traditionally had been validated against success in training rather than on-the-job performance was potentially problematic. Supporting studies regarding the relationship between recruit quality and military performance lacked persuasive power because proxy measures (e.g., attrition, promotion rates, or reenlistment eligibility) were used rather than actual measures of job performance. Congressional scrutiny of the ASVAB misnorming and surrounding issues of recruit quality and entry standards led to the Joint-Service Job Performance Measurement/Enlistment Standards Project – hereafter referred to as the JPM Project.

In July 1980, Robert B. Pirie, Assistant Secretary of Defense for Manpower, Reserve Affairs, and Logistics, directed the Military Services to establish a research and development program to link enlistment standards to job performance. As a result, the JPM Project was initiated. During the Fiscal Year 1983 budget hearings, this subject was again addressed by Congress and endorsed by the House Committee on Appropriations. The committee tasked the Office of the Secretary of Defense to provide direct oversight for Joint-Service research activities addressing job performance measurement and their eventual linkage to enlistment standards. Project A was the Army's contribution to JPM.

The Army viewed the Congressional directive as an opportunity to address a much larger set of personnel research questions. Could other selection and classification measures be developed to supplement the ASVAB? Could selection tests identify individuals more likely to complete their tour of service? Given the declining manpower pool, could tests be designed to achieve better classification? In addition to the JPM mandate, Project A resulted from the Army's need to address some very real policy issues and improve the design and functioning of the Army's selection/classification decision procedures.

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) organized the Manpower and Personnel Research laboratory to be responsible for this large-scale personnel research program. After several months of writing and rewriting, the Request for Proposals was released in the fall of 1981. In September 1982, a contract for Project A was signed with the Human Resources Research Organization (HumRRO) and its subcontractors, the American Institutes for Research, and Personnel Decisions Research Institute, Inc.

ARI's effort was designed to address a broad set of selection and classification issues using a very large, but integrated, database. Working from a single overarching design, the research program incorporated two sequential projects, **Project A** (1982–1989) and **Career Force** (1990–1994). Project A covered all initial instrument development work and all data collections, which involved the assessment of training performance and job performance during the first tour of duty for enlisted personnel in the U.S. Army. The Career Force project involved the assessment and prediction of job performance during the second tour of duty, that is, after soldiers reenlist and begin to take on supervisory responsibilities as a junior noncommissioned officer (NCO).

Collectively, these projects attempted to evaluate the selection validity and classification efficiency of different kinds of prediction information for different selection and classification goals (e.g., maximize future performance, minimize turnover/attrition) using a variety of alternative decision rules (i.e., "models"). Tackling such an ambitious objective required the development of a comprehensive battery of new tests and inventories, the creation of a wide variety of training and job performance measures for each job in the sample, four major worldwide data collections involving thousands of Army enlisted job incumbents for one to two days each, and the design and maintenance of the resulting database.

The most challenging aspect of the project was the need to develop a consensus among all the participants regarding hundreds of choices among measurement procedures, analysis methods, and data collection design strategies. Although decisions on these issues were made in the original design stage, many more occurred continuously as the projects moved

forward, driven by inflexible target dates for the major data collections. The fact that all major parts of the projects were completed within the prescribed time frames and according to the specified research design was a colossal achievement for all who participated.

The Political Environment of JPM

Linking Recruit Quality to Job Performance

The Department of Defense is the world's largest employer of young people. Depending on personnel requirements, the Department screens hundreds of thousands of youth for enlistment annually (Eitelberg, Laurence, & Waters, 1984). During the late 1970s, DoD screened approximately one million applicants each year, while that number declined to about 500,000 during the first years of the 21st century. The military's task in screening potential recruits is complicated by the fact that the available personnel pool is composed predominantly of young men and women who have never held a permanent full-time job. Thus screening is not based on actual military performance, which is not observable until after a recruit enlists, but on the statistical relationship between characteristics (e.g., aptitude, educational achievement, physical fitness, police record, etc.) assessed at the enlistment point and expected performance.[33]

Standards for entry into the Services are justified in terms of force readiness requirements and are influenced by the costs of recruiting qualified individuals for enlistment. When recruiting prospers, the Services raise their enlistment standards. When times are bad, the Services lower their standards and access somewhat lower recruit quality in order to meet their recruiting

[33] See Chapter 3 for a complete description of current screening procedures.

goals. The motivation to conduct research into the linkage between recruit quality and performance outcomes has often centered on budgetary matters. Military recruiting, assignment, and training of young unskilled people is an investment; the underlying purpose of the personnel selection and job classification process is to reduce the risk that an investment will be made in persons who are unable (or unwilling) to perform their duties. In 1997, the U.S. General Accounting Office estimated that it cost the Department of Defense about $35,000 to replace (recruit, train, and equip) each individual who fails to successfully complete a first tour of duty (U.S. General Accounting Office, 1997). As noted in Chapter 2, a more recent estimate put the cost to reach the soldier's first duty station at $50,000 (J. Thomas, personal communication, February 8, 2005). On the other hand, there are also costs associated with recruit quality levels; it is more difficult and costly to recruit high-quality youth (high school graduates with above average aptitude) than their lower-quality peers. Thus, recruit quality standards directly influence recruiting resource requirements (Sellman, 1999).

Resource Implications of Enlistment Standards

Under the all-volunteer force concept, the Services seek to encourage young people to consider and then, if qualified, choose enlistment in the military as an opportunity that matches their interests and plans. As mentioned earlier, resources required to support the enlistment process are directly tied to the enlistment standards set by the Services. Costs rise as recruit quality goals rise. Included are costs associated with maintaining a recruiting force to reach high-quality youth, providing information to the public through recruitment advertising, and funding programs which attract capable young

people (e.g., enlistment bonuses, education benefits). Because higher quality individuals are more expensive to recruit than lower quality people, decisions to set recruit quality goals at a given level necessarily incorporate recruiting resource costs.

From the beginning, the JPM Project was intended to inform the standards setting process by permitting the inclusion of performance as an enlistment criterion. The research goal was the development of a modeling capability that linked recruit quality, recruiting resources, and military job performance. Department of Defense and Service personnel planners and recruiting analysts could use the known relationship between recruit quality and predicted job performance to estimate the enlistment standards necessary to produce a desired level of performance, and to program and justify the resources required to obtain that performance within the personnel selection and job classification system (Green & Mavor, 1994; Wigdor & Green, 1986).

The Enabling of Project A/Career Force

Classic Validity Model

In addition to the coming together of a broad consensus about the need for this major research and development project, scientific advances had emerged which made an effort of this scale possible. The overall design of the Project A/Career Force program was intended to be fundamentally different from the conventional paradigm that dominated from 1906 to 1982, which consisted of computing the correlation between a single predictor score, or a predictor composite score, and a single criterion measure of performance obtained from a sample of job incumbents. Literally thousands of such estimates have been generated over the past 100 or more years (e.g., Ghiselli, 1973; Hunter & Hunter, 1984; Nathan & Alexander, 1988; Schmidt,

1988; Schmidt & Hunter, 1998; Schmitt, Gooding, Noe, & Kirsch, 1984).

There are several reasons why single investigators working to generate one bivariate distribution at a time has historically served as the dominant paradigm in most personnel research. In general, this resulted from the constraints imposed by technology, the structural characteristics of the research enterprise itself, and because of the contingencies built into the reward structures for individual investigators. When the basic outline of Project A/Career Force was proposed, there was no single organization or university group that had the resources necessary to carry it out. Coalitions of organizations had to form. There were also no practical means available for coordinating the efforts of researchers who are geographically scattered. Neither was there a technology for building a central database that could be accessed efficiently from remote locations.

There are exceptions to the above depiction of this dominant paradigm. Two of the more prominent ones are the AT&T management Progress Study (Bray, Campbell, & Grant, 1974) and the Sears executive selection studies (Bentz, 1968). Both were programmatic in nature, were coordinated efforts over relatively long periods of time, and dealt with the prediction of success in a broad class of "management" jobs that varied by function and by level in the organization. The AT&T study is particularly noteworthy because it was a "blind" longitudinal study (over 20 years) that did not use predictor information (scores from an assessment center) to make selection or promotion decisions. However, even with regard to these major exceptions, the number of different jobs was relatively circumscribed and the total sample sizes were still relatively

small. For example, the AT&T management progress study began with a total sample size of 550 new college hires. In contrast, the longitudinal validation component of Project A/Career Force began with a sample of almost 50,000 new recruits.

Military personnel research, and research sponsored by the military, is an important segment of the total personnel research record during the 20th century. The work of military psychologists during World War II is reasonably well-known and well-documented. Of special relevance to Project A and Career Force is the work on differential prediction and classification models by Brogden (1946a, 1951) and Horst (1955); on utility conceptions of validity by Brogden (1946b) and Brogden and Taylor (1950a); on the "structure of intellect" by Guilford (1957); on the establishment of critical job requirements by Flanagan and associates (Flanagan, 1954); and on the decision-theoretic formulations of selection and classification developed by Cronbach and Gleser (1965) for the Office of Naval Research. The last of these (*Psychological Tests and Personnel Decisions*) was hailed quite appropriately as a breakthrough in selection and classification; although the authors were the first to acknowledge the relevance of the initial work of Brogden and Horst. It was the culmination of a lengthy sequence of development. However, as impressive and as important as the military-sponsored research had been, it did not represent a major departure from the dominant paradigm described above.

Forging a New Path

Fortunately, along with the Army's need to address enlisted selection and classification as a system there were concomitant developments in the structure and technology of the personnel

research enterprise. For example, advances in computerized database management and electronic communication made it possible to design, create, edit, update, and maintain a very large database that could be accessed from anywhere in a very efficient manner. What is routine now was new and liberating in 1982.

Advances in computerization also permitted use of new testing technologies and the development of powerful, linear programming algorithms that made the estimation of classification efficiency and the comparison of alternative selection/classification strategies using the entire Army database a very manageable analytic problem. Certainly, the development of confirmatory techniques within the general domain of multivariate analysis models opened up a number of powerful strategies for generalizing research findings from a sample of jobs to the entire population of jobs in the organization's personnel system.

Finally, I/O psychologists realized in the 1970s that their fundamental task was to learn things about an appropriately defined population, and not to learn more and more specific things about specific samples. This realization changed the field's approach to the estimation of selection validity and classification efficiency. Meta-analysis and corrections for attenuation and restriction of range were no longer "risky" games to play. They were a conservative and necessary part of statistical estimation. These technological and academic developments provided a foundation for the work done on Project A/Career Force.

Project A/Career Force constitutes the largest, single research effort in the history of personnel research, by some orders of magnitude. There were a number of very large substantive

pieces to the overall design, each of which was the concern of several investigators under the direction of a single task leader. The separate pieces were interdependent and had to come together on a specific date (perhaps several years in the future), such that a particular phase of the data collection could begin. The data collection dates were set far in advance and were driven by the requirement to assess a specific cohort of new recruits as it was inducted, finished training, moved on to the job, and then either left the Army or reenlisted. That is, once the overall project started, it could not stop or deviate to any significant degree from the agreed upon data collection schedule, which spanned an 8-year period. There was zero tolerance for failure, and the projects had to be managed with this reality in mind. This level of "managed collaboration" across four independent organizations was a new but necessary development for conducting personnel research.

In summary, in 1982, those working on Project A hoped they had designed a research program that would bear directly on the major policy and design parameters of the selection/classification decision process such that the results would be directly useful for meeting the system's needs, both as they existed initially and as changes took place. Simultaneously, they hoped that by considering an entire system and population of jobs at once, and by developing measures from a theoretical/taxonomic base, the science of industrial and organizational psychology would also be served.

Research and Development Strategy

The JPM Project comprised three phases to be implemented by the Services: (a) Determine the feasibility of measuring hands-on job performance; (b) if feasible, validate ASVAB against on-the-job performance; and (c) develop an enlistment standards

cost/performance trade off model that linked recruit quality, recruiting resources, and job performance. The overall project strategy called for each Service to develop and demonstrate a variety of measurement approaches which could be used to link enlistment standards to job performance.

Project A Research Objectives

In compliance with the JPM research strategy, Project A's objectives spanned a continuum from operational/applied concerns to more theoretical interests and may be summarized as follows:

Predictor Measurement

- Identify the constructs that constitute the universe of information available for selection/classification into entry-level skilled jobs given no prior job experience on the part of the applicant.
- Develop predictor measures for the constructs identified as "best bets" for enhancing selection and classification.

Criterion Measurement

- Develop measures of entry-level job performance that can be used as criteria against which to validate selection/classification measures.
- Develop a general model of performance for entry-level skilled jobs.
- Develop a complete array of valid and reliable measures of second-tour performance as an Army NCO, including its leadership/supervision aspects.

- Develop a model of NCO performance that identifies the major components of second-tour job performance.

Validation

- Validate existing selection measures (i.e., ASVAB) against training and job performance criterion measures.
- Based on "best bet" constructs, validate a battery of new selection and classification measures.
- Carry out a complete incremental predictive validation of (a) the ASVAB and an experimental battery of predictors, (b) measures of training success, and (c) the full array of first-tour performance criteria using the second-tour job performance measures as criteria.
- Estimate the degree of differential prediction across (a) major domains of predictor information (e.g., abilities, personality, interests), (b) major factors of job performance, and (c) different types of jobs.
- Determine the extent of differential prediction across racial and gender groups for a systematic sample of individual differences, performance factors, and jobs.
- Develop the analytic framework needed to evaluate the optimal prediction equations under varying conditions.

Other Research Objectives

- Develop a utility scale for different performance levels across jobs.

- Design and develop a fully functional and user-friendly research database that includes all relevant personnel data on the three cohorts of new Army accessions included in the research program.

Overall Research Design

Project A researchers spent the first 6 months planning, documenting, reviewing, modifying, and redrafting research plans, requests for participants/subjects, administrative support requests, and budgetary plans. They also began the comprehensive literature reviews and job analyses. The final detailed version of the operative research plan was published as ARI Research Report No. 1332, *Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Project A Research Plan.*

Sampling Jobs

In 1982 the population of enlisted jobs included approximately 275 different Military Occupational Specialties (MOS). Because data could not be collected from all of them, MOS were sampled representatively. This meant there would be a trade-off in the allocation of research resources between the number of jobs researched and the number of incumbents sampled from each job. Cost considerations dictated that 18 to 20 MOS could be studied if the initial goal was 500 incumbents per job, and this assumed that a full array of job-specific performance measures would be developed for only a subset of those MOS.

An initial sample was drawn on the basis of the following considerations:

1. High-density jobs that would provide sufficient sample sizes for statistically reliable estimates of validity.

2. Representation of the Army's designated Career Management Fields (CMF), which are clusters of related jobs.
3. Representation of the jobs most crucial to the Army's missions.

The initial set of 19 MOS included only 5% of Army jobs, but represented 44% of the soldiers recruited in FY81. Similarly, of the total number of women in the Army, 44% were represented in the sample. A supplemental cluster analysis of MOS (based on task content similarity) was carried out via SME judgment to evaluate and refine the sample. Table 8.1 shows the resulting MOS ($N = 21$) that were studied over the course of the Project A/Career Force research program. "Batch A" MOS received the most attention in that soldiers in these jobs were administered a full array of first- and second-tour job performance measures, including hands-on work sample tests, written job knowledge tests, and Army-wide and MOS-specific ratings. soldiers in "Batch Z" were not measured as extensively with regard to the job performance criterion measures.

Table 8.1. Military Occupational Specialties (MOS) that were Sampled

MOS Batch A		MOS Batch Z	
11B	Infantryman	12B	Combat Engineer
13B	Cannon Crewmember	16S	MANPADS Crewman
19E/K	Army Tank Crewman	27E	Tow/Dragon Repairer
31C	Single Channel Radio Operator	29E	Comm-electronics Radio Repairer
63B	Light-Wheel Vehicle Mechanic	51B	Carpentry/Masonry Specialist
71L	Administrative Specialist	54B	NBC Specialist
88M	Motor Transport Operator	55B	Ammunition Specialist
91A/B	Medical Specialist/Medical NCO	67N	Utility Helicopter Repairer
95B	Military Police	76Y	Unit Supply Specialist
		94B	Food Service Specialist
		96B	Intelligence Analyst

Data Collection Design

The basic design framework and major samples encompassed two major cohorts, each of which was followed into their second tour of duty and which collectively produced six major research samples. Development of the predictor and criterion measures administered during the major phases of this research involved dozens of additional smaller data collection efforts as well, for purposes of pilot and field testing. Each of the six major data collections is briefly characterized below.

Concurrent validation (CVI) sample. This sample was drawn from soldiers who entered the Army between July 1983 and June 1984 and had been in the Army for 18 to 24 months. Data were collected from these soldiers and their supervisors at 13 posts in the continental United States and at multiple locations in Germany. Batch A soldiers (see Table 8.1) were assessed for 1½ days on the first-tour job performance measures and for ½ day on the new predictor measures (the Trial Battery). Batch Z soldiers were tested for ½ day on a subset of the performance measures and ½ day on the Trial Battery.

Longitudinal validation predictor (LVP) sample. Virtually all new recruits who entered the Army into one of the sampled MOS from August 1986 through November 1987 were assessed on the 4-hour Experimental Battery (to be described) within 2 days of first arriving at their assigned reception battalion.

Longitudinal validation end-of-training (LVT) sample. End-of-training performance measures were administered to those individuals in the LVP sample who completed advanced individual training (AIT), which could take from 2 months to 6 months, depending on the MOS.

Longitudinal validation first-tour (LVI) sample. The individuals in the 86/87 cohort who were measured with the Experimental Battery, completed training, and remained in the Army were assessed with the first-tour job performance measures when they had roughly between 18 and 24 months of service. Data collections were conducted at 13 posts in the U.S, and multiple locations in Europe.

Concurrent validation second-tour (CVII) sample. The same data collection teams that administered the first-tour performance measures to the LVI sample administered second-tour performance measures at the same location and during the same

time periods to a sample of junior NCOs from the 83/84 cohort who were in their second tour of duty (4 to 5 years of service).

Longitudinal validation second-tour (LVII) sample. This sample includes members of the 86/87 cohort from the Batch A MOS who were part of the LVP (predictors), LVT (training performance measures), and LVI (first-tour job performance measures) samples and who reenlisted for a second tour. The revised second-tour performance measures were administered at 15 U.S. posts, multiple locations in Germany, and two locations in Korea.

Research Instrument Development: Predictors

A major objective of Project A was to develop an experimental battery of new tests that had maximum potential for enhancing selection and classification decisions for the entire enlisted personnel system. Rather than the traditional approach of basing the selection of predictor constructs on a job analysis of the occupations in question, the general strategy was to identify a universe of potential predictor constructs for the population of enlisted MOS and then sample appropriately from it. The next steps were to develop measures for each sampled construct when that construct was above some threshold of criticality, and refine and improve the measures through a series of pilot and field tests. The intent was to develop a predictor battery that was maximally useful for selection and classification into an entire population of jobs, and that provided maximal incremental information beyond that provided by the ASVAB.

Critical Constructs

After a thorough literature review and several iterations of consolidation and revision, the research team identified a list of 53 potentially useful predictor variables. A sample of 35

personnel selection experts was then asked to estimate the expected correlations between each predictor construct and an array of potential performance factors. The estimates were analyzed and compared to meta-analytic information from the empirical literature.

All the available information was then used to arrive at a final set of variables for which new measures would be constructed. The full array of candidate variables, rank ordered within domain is shown in Table 8.2.

Table 8.2. "Best Bet" Predictor Variables Rank Ordered, Within Domain, by Priority for Measure Development

Cognitive Ability Variables	*Psychomotor Abilities*
1. Spatial Visualization/Rotation	1. Multilimb Coordination
2. Spatial Visualization/Field Independence	2. Control Precision
3. Spatial Organization	3. Manual Dexterity (later replaced by Movement Judgment
4. Reaction Time	
5. Induction (reasoning)	
6. Perceptual Speed & Accuracy	
7. Numerical Ability	
8. Memory	

Biodata/Temperament Variables	*Interest Variables*
1. Adjustment	1. Realistic
2. Dependability	2. Investigative
3. Achievement	3. Conventional
4. Physical Condition	4. Social
5. Potency	5. Artistic
6. Locus of Control	6. Enterprising
7. Agreeableness/Likeability	
8. Validity Scales (not viewed as a predictor, per se, but a necessary component of this type of measure)	

Experimental Predictor Tests

The development of new tests to assess the high priority constructs went through several stages, including a large-scale field test (Preliminary Battery sample) that administered the new experimental tests along with a set of established marker tests for several of the constructs (see Peterson, et al. 2001).

Included in these efforts were the (a) development of the software for a computerized battery of perceptual/psychomotor tests, (b) the design and construction of a special response pedestal permitting a variety of responses (e.g. one-hand tracking, two-hand coordination), and (c) the acquisition of portable computerized testing stations. Several paper-and-pencil cognitive tests and two inventories were also developed. One inventory assessed relevant vocational interests and the other focused on major dimensions of personality and biographical history. Finally, based on the extensive literature on job outcomes provided by studies of job satisfaction and work motivation, an inventory was developed that asked the respondent to reflect the strength of his or her preferences for certain job outcomes (e.g. rewards) on a 7-point scale. The final form of this inventory was titled the Job Orientation Blank (JOB), and it included 29 items. Factor analyses of the field test data suggested that the 29 items could be grouped into six factors: Job Security, Job Status, Serving Others, Autonomy, Routine Work, and Ambition/Achievement.

Basic Predictor Composite Scores

The ASVAB together with the experimental tests produced a set of 72 scores. This number was too large for validation analyses that take advantage of idiosyncratic sample characteristics, (e.g. multiple regression). Therefore, a series of analyses was conducted to determine a smaller set of predictor composite scores that would preserve the heterogeneity of the full set of basic scores to the greatest extent possible. These analyses included exploratory factor analyses and confirmatory factor analyses guided by considerable prior theory and empirical evidence (McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Peterson et al, 1990). A final set of 31 composites was identified and is shown in

Table 8.3. Collectively, the Experimental Battery and the ASVAB were intended to be a comprehensive and representative sample of predictor measures from the population of individual differences that are relevant for personnel selection and classification, and which can be measured in a standardized fashion at the time of entry.

Job Analyses and Criterion Development

In contrast to the predictors, virtually all criterion development in Project A/Career Force was based on extensive job analyses. Task descriptions, critical incident analysis, and interviews with SMEs were used extensively. Relevant job manuals and available Army Occupational Survey results were used to enumerate the complete population of major tasks for each MOS (n = 100 – 150). These tasks were then grouped into clusters and rated for criticality and difficulty by panels of SMEs.

Additional panels of SMEs were used in a workshop format to generate approximately 700 to 800 critical incidents of effective and ineffective performance per MOS that were specific to each MOS, and approximately 1,100 critical incidents that could apply to any MOS. For both the MOS-specific and Army-wide critical incidents, a retranslation procedure was carried out to establish dimensions of performance.

Together, the task descriptions and critical incident analyses of MOS-specific and Army-wide performance were intended to produce a detailed content description of the major components of performance in each MOS. These are the job analysis results that were used to begin development of the performance criterion measures.

Table 8.3. Experimental Battery Composite Scores and Constituent Basic Scores

ASVAB Composites	*Computer-Administered Test Composites*
Quantitative	Psychomotor
Math Knowledge	Target Tracking 1 Distance
Arithmetic Reasoning	Target Tracking 2 Distance
Technical	Cannon Shot Time Score
Auto Shop	Target Shoot Distance
Mechanical Comprehension	Movement Time
Electronics Information	Pooled Movement Time
Speed	Perceptual Speed
Coding Speed	Perceptual Speed & Accuracy (DT)
Number Operations	Target Identification (DT)
Verbal	Basic Speed
Word Knowledge	Simple Reaction Time (DT)
Paragraph Comprehension	Target Identification (DT)
General Science	Perceptual Accuracy
Spatial Test	Perceptual Speed & Accuracy (PC)
Assembling Objects	Target Identification (PC)
Object Rotation	Basic Accuracy
Mazes	Simple Reaction Time (PC)
Orientation	Choice Reaction Time (PC)
Maps	Number Speed and Accuracy
Reasoning	Number Speed (Operation DT)
	Number Memory (PC)
	Short Term Memory
	Short Term Memory (PC)
	Short Term Memory (DJ)

DT = Decision Time
PC = Proportion Correct

Table 8.3 (continued). Experimental Battery Composite Scores and Constituent Composite Scores

ABLE Composites	*AVOICE Composites*
Achievement Orientation	Rugged Outdoors
Self-Esteem	Combat
Work Orientation	Rugged Individualism
Energy Level	Firearm Enthusiast
Leadership Potential Domain	Audiovisual Arts
Dependability	Drafting
Traditional Values	Audiographics
Conscientiousness	Aesthetics
Non-delinquency	Interpersonal
Adjustment	Medical Service
Emotional Stability	Leadership Guidance
Cooperativeness	Skilled/Technical
Cooperativeness	Science/Chemical
Internal Control	Computers
Internal Control	Mathematics
Physical Condition	Electronic Communication
Physical Condition	Administrative
	Clerical/Administrative
JOB Composites	Warehouse/Shipping
High Job Expectations	Food Service
Pride	Food Service—Professional
Job Security	Food Service—Employee
Serving Others	Protective Service
Ambition	Fire Protection
Job Routine	Law Enforcement
Routine	Structural/Machine
Job Autonomy	Mechanics
Autonomy	Heavy Construction
	Electronics
	Vehicle Operation

The job analysis goals for the second tour included the description of the major differences in technical task content between the first and second tour and the description of the leadership/supervision components of the junior NCO position. The task analysis and critical incident steps used for the first tour were also used for the second tour. In addition, a special

46-item job analysis instrument, the Supervisory Description Questionnaire, was constructed and used to collect item criticality judgments from SMEs. Consequently, the supervisory/leadership tasks judged to be critical for an MOS became part of the population of tasks for that MOS.

Performance Criteria

The general goals of training performance and job performance measurement were to define, or model, the total domain of performance in some reasonable way and then develop reliable and valid measures of each major factor. The general procedure for criterion development followed a basic cycle of a comprehensive literature review, initial instrument construction based on the job analyses described above, pilot testing, instrument revision, field testing, and proponent (i.e. management) review. The specific measurement goals were to:

1. Develop standardized measures of training achievement for the purpose of determining the relationship between training performance and job performance.
2. Make a state-of-the-art attempt to develop job sample or "hands-on" measures of job task proficiency.
3. Develop written proceduralized knowledge measures of job task proficiency.
4. Develop rating scale measures of performance factors that are common to all first-tour enlisted MOS (Army-wide measures), as well as for factors that are specific to each MOS.
5. Compare hands-on measurement to paper-and-pencil tests and rating measures of proficiency on the same tasks (i.e. a multitrait, multimethod approach).

6. Evaluate existing archival and administrative records as possible indicators of job performance.

The Initial Theory

Criterion development efforts were guided by a model that viewed performance as truly multidimensional. For the population of Army entry-level enlisted positions there were two major types of performance components: (a) those that reflect specific technical tasks or specific job behaviors that are not required for other jobs, and (b) components that are defined and measured in the same way for every job (i.e. Army–wide) such as contributions to teamwork, continual self-development, support for the norms and customs of the organization, and perseverance. The working model of total performance with which Project A began viewed performance as multidimensional within these two broad categories of performance requirements.

Training Performance Measures

Because a major program objective was to determine the relationships between training performance and job performance and their differential predictability, if any, a comprehensive training achievement test was constructed for each MOS. The content of the program of instruction (POI) was compared to the previously determined content of the critical job tasks for each MOS. For the POI content judged to be reflective of critical job tasks, items were written to represent proceduralized knowledge reflective of how to do a task. After pilot testing, revision, field testing, and Army proponent review, the result was a 150 to 200 item training achievement test for each MOS included in the research samples. Rating scales were also developed for completion by peers and drill instructors at

the end-of-training (EOT). The following six scores were obtained from the EOT measures.

From the EOT Achievement Test

1. Technical content score (TECH).
2. Army-wide basic training content total score (BASC).

From the EOT Rating Scales

3. Technical achievement and effort (ETS)
4. Maintaining personal discipline (MPD).
5. Physical fitness and military bearing (PFB).
6. Leadership potential (LDR)

First Tour (entry level) Performance Measures

Criterion development for assessment of first-tour performance proceeded from the two basic types of job analysis information, task analysis and critical incident analysis. The task-based information was used to develop standardized hands-on job samples, paper-and-pencil job knowledge tests, and rating scales for each Batch A MOS. These measures were intended to assess knowledge and proficiency on the critical tasks associated with each MOS. Roughly 30 tasks per MOS were covered by the written job knowledge tests and rating scales, and about one-half of those tasks were also tested using a hands-on format. For the hands-on simulations, each examinee passed through a testing station for each of the 15 ($\pm$ 2) major job tasks and was asked to perform a standardized simulation of the task, using real equipment if possible. Each measure went through multiple rounds

of pilot testing and revision before being used for validation purposes.

Given the critical incident analyses, a modified behaviorally-anchored rating scale procedure was used to construct six to nine rating scales for performance factors specific to a particular job and also for ten performance factors that were defined in the same way and relevant for all jobs. The critical incident procedure was also used with workshops of combat veterans to develop rating scales of expected combat effectiveness. Ratings were gathered from both peers and supervisors of first-tour soldiers. Rating scale development activities included the creation of procedures for identifying qualified raters and a comprehensive rater training program.

The final category of job performance criterion measure was produced by a search of the Army's archival records for potential performance indicators. First, all possibilities were enumerated from the major sources of such records maintained by the Army. Considerable exploration of these sources identified the most promising indexes, which were then investigated further to determine their usefulness as criterion measures.

The result of the confirmatory factor analyses described earlier was a five factor model of first-term (entry level) performance that was very robust across samples. The definition of the factors and the basic criterion scores that comprise them are described below.

Five Factor Model of First-Tour Job Performance

1. Core Technical Proficiency (CTP) represents the proficiency with which the Soldier performs the tasks that are "central" to the MOS. These tasks represent the core of

the job, and are its primary definers. This performance construct does not include the individual's willingness to perform the task or the degree to which the individual can coordinate efforts with others. It refers to how well the individual can execute the core technical tasks the job requires, given a willingness to do so. CTP was measured through:

   - Hands-On Test – MOS Specific Tasks
   - Job Knowledge Test – MOS-Specific Tasks

2. General Soldiering Proficiency (GSP). In addition to the core technical content specific to an MOS, individuals in every MOS are also responsible for being able to perform a variety of general Soldiering tasks – for example, "determines a magnetic azimuth using a compass; recognizes and identifies friendly and threat vehicles." Performance on this construct represents overall proficiency on these general Soldiering tasks. Again, it refers to how well the individual can execute general Soldiering tasks, given a willingness to do so. GSP was measured through:

   - Hands-On Test – Common Tasks
   - Job Knowledge Test – Common Tasks

3. Effort and Leadership (ELS) is a performance construct that reflects the degree to which the individual exerts effort over the full range of job tasks, perseveres under adverse or dangerous conditions, and demonstrates leadership and support toward peers. That is, can the individual be counted on to carry out assigned tasks, even under adverse conditions, to exercise good judgment, and to be generally dependable and proficient? While appropriate knowledge and skills are necessary for successful performance, this construct is meant only to reflect

the individual's willingness to do the job required and to be cooperative and supportive with other soldiers. ELS was assessed through:

- Admin: Number of Awards and Certificates
- Army-Wide Rating Scales: Overall Effectiveness Rating Scale
- Army-Wide Rating Scales: Effort/Leadership Ratings Factor
- Average of MOS Specific Ratings Scales

4. Maintaining Personal Discipline (MPD). Reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self-control, demonstrates integrity in day-to-day behavior, and does not create disciplinary problems. People who rank high on this construct show a commitment to high standards of personal conduct. MPD was measured through:
   - Admin: Number of Disciplinary Actions
   - Admin: Promotion Rate Score
   - Army-Wide Rating Scales: Personal Discipline Ratings Factor
5. Physical Fitness and Military Bearing (PFB). Represents the degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition. PFB was assessed through:
   - Admin: Index – Physical Readiness Score
   - Army-Wide Rating Scales: Physical Fitness/Bearing Ratings Factor

Note that the first two factors are represented by the Hands-On work sample test and the job knowledge tests while the last three factors are each represented by both rating scales and archival administrative measures. Again, this factor solution represented the best fitting a priori model in both the concurrent and longitudinal cohort samples and cross validated from one cohort to the other with no loss in the accuracy of fit. It was a very stable representation.

Second Tour (NCO) Measures

The goal of criterion measurement of second-tour job incumbents was to provide a comprehensive assessment of junior NCO performance. While the job analyses of the second-tour job indicated that there is considerable overlap between first-and second-tour performance requirements, almost all of the overlap occurs in the technical task content of the position, although NCOs are expected to perform at somewhat higher levels on the technical tasks. Consequently, with relatively few modifications, the first-tour technical performance measures were used to measure second-tour performance. The differences occur because soldiers begin to take on leadership responsibilities during the second tour which are substantial and critical.

The second-tour job analysis results identified six additional MOS-specific leadership dimensions and three Army-wide leadership dimensions. These findings led to the development of several additional rating scales for measuring supervisory performance. (The second-tour rating scale measures of the non-leadership/supervisory performance factors were very similar to the rating scales for the first tour). A set of supervisory performance rating scales was created to measure the following dimensions: acting as a role model,

communication, personal counseling, monitoring subordinate performance, organizing missions/operations, personnel administration, and performance counseling/correcting. Also, the attempts to collect peer ratings in CVII met with only limited success. As a result, only supervisor ratings were collected in LVII.

New Second-Tour Measurement Methods

Based on a review of the literature and consideration of feasibility, two additional methods were developed for assessing NCO performance. The first was a set of assessment center-like, role-play exercises, and the second was a written situational judgment test (SJT).

Supervisory Role-Play Exercises. Role-play exercises were developed to simulate three of the critical and distinct NCO supervisory tasks.

- Counseling a subordinate with personal problems that affect performance.
- Counseling a subordinate with a disciplinary problem.
- Conducting one-on-one remedial training.

Information for the development of the supervisory simulations was drawn from a number of sources, including Army NCO training materials and the second-tour job analyses. The format for the simulations was for the examinee to play the role of a supervisor. The subordinate was played by a trained confederate who also scored the performance of the examinee on a number of specific dimensions within the three major tasks.

Situational Judgment Tests. The purpose of the SJT was to evaluate the quality of judgments about how to most effectively react in typical supervisory problem situations. A critical

incident methodology was used to generate situations for inclusion in the SJT. Response options were developed through a combination of input from pilot test SMEs and examinees from the field tests. Independent groups of SMEs scaled the response options in terms of desirability of using each option to address the problems described in the item (i.e. what should be done).

A Model of Second-Tour Performance

A confirmatory factor analysis procedure similar to that used for the first tour analysis yielded a six factor model of NCO performance. The sixth factor represents the leadership supervisory component. The other five factors are very similar in content to the first term model. The six factors and the basic criterion scores that comprise them are shown below.

1. Core Technical Proficiency (CTP)
    - Hands-On Test – MOS Specific Tasks
    - Job Knowledge Test – MOS Specific Tasks
2. General Soldiering Proficiency (GSP)
    - Hands-On Test – Common Tasks
    - Job Knowledge Test – Common Tasks
3. Achievement and Effort (AE)
    - Admin: Number of Awards and Certificates
    - Army-Wide Rating Scales: Overall Effectiveness Rating Scale
    - Army-Wide Rating Scales: Technical Skill/Effort Ratings Factor
    - Average of MOS Specific Rating Scales

- Average of Combat Prediction Rating Scales

4. Personal Discipline
    - Admin: Number of Disciplinary Actions
    - Army-Wide Rating Scales: Personal Discipline Ratings Factor
5. Physical Fitness and Military Bearing (PFB)
    - Admin: Physical Readiness Score
    - Army-Wide Ratings Scales: Physical Fitness/Bearing Ratings Factor
6. Leadership (LDR)
    - Admin: Promotion Rate Score
    - Army-Wide Rating Scales: Leading/Supervising Ratings Factor
    - Individual scores from each of the three role plays
    - Situational Judgment Test – Total Score

Results Summary

Overall, the validation results indicated that:

- The ASVAB is a powerful predictor of performance, be it in training, during the first tour of enlistment, or in the second tour; and it has a substantial correlation with all the major components of performance. Its validity for performance prediction is not dependent on the criterion measurement method.
- Many tests in the Experimental Battery showed high predictive validities as well. Incremental validities over ASVAB were limited to predictors of the non-

technical task components of performance, when averaged over MOS. However, a number of the spatial and psychomotor tests showed incremental validities for the prediction of performance on critical core tasks in specific MOS (Walker & Rumsey, 2001)

- Current performance does predict future performance and with considerable convergent/divergent validity across performance factors.
- While the use of the differential prediction available in the ASVAB subtests alone can yield significant classification efficiency (Zeidner, Johnson, & Scholarios, 1997), the addition of the Experimental Battery increased aggregate performance (classification efficiency) even more (Rosse, Peterson, & Campbell, 2001).

Development of the DoD Recruit Quality Benchmarks

One of the major objectives of the JPM Project was development of a mathematical model to link recruit quality, recruiting resources, and job performance. Working with the National Research Council, the Department of Defense used that model to establish the DoD recruit quality benchmarks in 1991 (Sellman, 1997). In general, enlistment standards are based on judgments by Service policymakers as to the level of job performance required. However, standards should be guided by empirical evidence of the relationship between recruit quality and the required level of performance. It is extremely difficult to specify an absolute value of performance that can be considered sufficient to guarantee successful military mission accomplishment. Even so, the research performed within the

JPM Project developed reliable and valid measures of individual job performance which became the basis for the linkage model.

How do DoD and the Services decide how many high school diploma graduate and above average aptitude recruits to enlist? The goal is to maximize recruit quality (aptitude and education) while minimizing recruiting, training, and attrition costs. The linkage model specifies the number of high-quality recruits who will provide the desired level of job performance for the least cost (Harris, et al., 1991; McCloy, 1994; Smith & Hogan, 1994; Wise, 1994). Scores on the hands-on performance tests define the job performance variable (Green & Mavor, 1994; Wigdor & Green, 1991). Costs reflect resources needed for training, compensation, and recruiting (e.g., recruiter compensation and money for advertising, education benefits, and enlistment bonuses). Using these relationships, the model allows "what-if" analyses to examine how changes in one of more of these variables affect the others. For example, the model can answer how decreasing the advertising budget or decreasing the number of recruiters, but increasing the money available for enlistment bonuses and education benefits, would affect recruit quality and job performance.

Recruit quality benchmarks are used to help ensure that recruit performance is sufficient to complete military missions. The linkage model cannot estimate how much quality is enough. Rather, that is a policy decision – personnel planners and recruiting analysts within DoD and the Services set the desired level of performance. Nevertheless, the model can help specify a cohort of recruits that will provide the desired level of performance for the lowest cost.

What should be the desired performance level? The performance level identified by the policy analyst is a minimally acceptable value. The Department has chosen the level of performance provided by the 1990-91 enlisted cohort, a group that produced satisfactory performance during Operations Desert Shield and Desert Storm. Specifying this level of desired performance resulted in recruit quality benchmarks that call for 60% of recruits to score above the 50th percentile on the AFQT and 90% to have high school diplomas (Sellman, 1994). For the most part, the Services have met or exceeded these benchmarks over the 25 years since the linkage model was implemented.

Contributions of the JPM Research

The Joint-Service JPM Project represents many years of intensive research by the Department of Defense and the Services to develop and demonstrate measures of on-the-job performance and to link those measures to enlistment standards. As the data presented in this chapter demonstrate, the Services developed reliable measures of job performance that closely replicated actual job performance on representative samples of job tasks essential to their occupations. The findings also revealed that higher quality recruits consistently displayed better hands-on job performance throughout the first term of service than did their lower scoring counterparts.

Thus, as a result of the JPM research, it was demonstrated that hands-on job performance could be measured and that ASVAB and enlistment standards could be validated against those measures. In addition, the job performance information was incorporated into a mathematical model that allows the Department of Defense to develop and defend recruiting budgets based on desired levels of recruit quality and job performance. With the implementation of the model, if

Congress asked what happens if the recruiting budget is cut by 10 percent, the Department can respond that recruit quality will go down by *X* percent and performance will drop by *Y* percent.

For years, industrial psychologists contended that job performance was the ultimate criterion for validating selection tests. In fact, S. Rains Wallace (1965), an eminent psychologist, once called it the holy grail of industrial psychology. Measuring job performance is a very expensive proposition. With the support of Congress and the DoD's effort to recover from the embarrassing misnorming episode, $40 million was made available for the JPM Project. Another aspect of this research effort that made it unique was its sustainability. It was widely recognized as a project of great merit, and it lasted for over 15 years, spanning five presidential administrations, both Democrat and Republican.

References

Bentz, V. J. (1968). The Sears experience in the investigation, description, and prediction of executive behavior. In J.A. Myers (Ed.), *Predicting managerial success*. Ann Arbor, MI: Foundation for Research on Human Behavior.

Bray, D. W., Campbell, R. J., & Grant, D. L. (1974). *Formative years in business.* New York: Wiley.

Brogden, H. E. (1946a). An approach to the problem of differential prediction. *Psychometrika, 11*, 139-154.

Brogden, H. E. (1946b). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology, 37*, 65-76.

Brogden, H. E. (1951). Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. *Educational and Psychological Measurement, 11,* 173-196.

Brogden, H. E., & Taylor, E. K. (1950a). The dollar criterion-applying the cost accounting concept to criterion construction. *Personnel Psychology, 3*, 133-154.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed). Urbana, IL: University of Illinois Press.

Eitelberg, M. J. (1988). *Manpower for military occupations.* Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel).

Eitelberg, M. J., Laurence, J. H., & Waters, B. K. (1984). *Screening for service: Aptitude and education criteria for military entry.* Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Installations and Logistics).

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26*, 461-477.

Green, B. F., & Mavor, A. S. (Eds.). (1994). *Modeling cost and performance for military enlistment.* Washington, DC: National Academy Press.

Guilford, J. P. (1957). *A revised structure of intelligence* (Report No. 19). University of Southern California Psychological Laboratory.

Harris, D. A., McCloy, R. A., Dempsey, J. R., Roth, C., Sackett, P. R., Hedges, L. V., Smith, D. A., & Hogan, P. F. (1991). *Determining the relationship between recruit characteristics and job performance: A methodology and a model* (FR-PRD-90-17). Alexandria, VA: Human Resources Research Organization.

Horst, P. (1955). A technique for the development of a multiple absolute prediction battery. *Psychological Monographs* (No. 390).

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 98*, 72-98.

McCloy, R. A. (1994). Predicting job performance scores for jobs without performance data. In B. F. Green & A. S. Mavor (Eds.), *Modeling cost and performance for military enlistment* (pp. 61-99). Washington, DC: National Academy Press.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43*, 335-354.

Meyer, E. C. (1980). *The hollow army* (White Paper). Washington, DC: Department of the Army.

Nathan, B. R., & Alexander, R. A. (1988). A comparison of criteria for test validation: A meta-analytic investigation. *Personnel Psychology, 41*, 517-536.

Peterson, N. G., Hough, L. M., Dunnette, M. D., Rosse, R. L., Houston, J. S., Toquam, J. L., & Wing, H. (1990). Project A: specification of the predictor domain and development of new selection/classification tests. *Personnel Psychology, 43*, 247-276.

Rosse, R. L., Campbell, J. P. & Peterson, N. G. (2001). Personnel classification and differential job assignments: Estimating classification gains. In J. Campbell & D. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 453-506). Mahwah, NJ: Erlbaum.

Schmidt, F. L. (1988). Validity generalization and the future of criterion related validity. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 173-189). Hillsdale, NJ: Erlbaum.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407-422.

Sellman, W. S. (1997). *Public policy implications for military entrance standards.* Keynote address presented at the 39th Annual Conference of the International Military Testing Association, Sydney, Australia.

Sellman, W. S. (1999). *Military recruiting: The ethics of science in a practical world.* Invited address to the Division of Military Psychology, 107th Annual Convention of the American Psychological Association, Boston, MA.

Smith, D. A. & Hogan, P. F. (1994). The accession quality cost/performance trade-off model. In B. F. Green & A. S. Mavor (Eds.), *Modeling cost and performance for military enlistment* (pp. 105-128). Washington, DC: National Academy Press.

U.S. General Accounting Office (1997, January). *Military attrition: DoD could save millions by better screening enlisted personnel* (NSIAD-97-39). Washington, DC: Author.

Walker, C. B., & Rumsey, M. G. (2001). Application of findings: ASVAB, new aptitude tests, and personnel classification. In J. Campbell & D. Knapp (Eds.) *Exploring the limits in personnel selection and classification* (pp. 457-466). Mahwah, NJ: Erlbaum.

Wallace, S. R. (1965). Criteria for what? *American Psychologist, 20,* 411-417.

Wigdor, A. K. & Green, B. F. (Eds.). (1986). *Assessing the performance of enlisted personnel.* Washington, DC: National Academy Press.

Wigdor, A. K. & Green, B. F. (1991). *Performance assessment in the workplace* (Vols. 1 & 2). Washington, DC: National Academy Press.

Wise, L. L. (1994). Setting performance goals for the DoD linkage model. In B. F. Green & A. S. Mavor (Eds.), *Modeling cost and performance for military enlistment* (pp. 37-60). Washington, DC: National Academy Press.

Zeidner, J., Johnson, C.D., & Scholarios, D. (1997). Evaluating military selection and classification systems in the multiple job context. *Military Psychology, 9,* 169-186.

CHAPTER 9

ARMY COMPETENCY TESTING: SKILL QUALIFICATION TESTS (SQT)

An Historical Study and Analysis of the Army Personnel Testing Program

1959-2008

Roy Campbell
Human Resources Research Organization (Retired)

The Beginning

Although the focus of this section is on the Army Skill Qualification Test (SQT), it is necessary to understand that the SQT was only one part of an overall revolution in Army training and performance doctrine. It is important to not only grasp the extent of these other developments, but to also have an appreciation of the tenor of the times and conditions under which they occurred. During the early 1970s there was a move toward change within the Army. As the Army involvement in Viet Nam began to draw down, younger career officers and emerging senior leadership looked back critically at their own wartime experiences in an unpopular and in an ultimately strategically unsuccessful effort. Those reflections, coupled with the American societal changes of the 1960s, the emerging Volunteer Army, and the changing threat capabilities, made many realize that the Army they had known had to change. While the Army produced many significant leaders from this group, the most influential single architect of what was to occur was General William E. DePuy.

General DePuy, who first entered the Army as an enlisted Soldier with the South Dakota Army National Guard, was commissioned out of South Dakota State University in 1941. He became a World War II battalion commander, rising to command the 1st Infantry Division in Viet Nam in 1966-67. During the early 1960s, then Major General DePuy, on a Pentagon assignment, had interacted closely with Secretary of Defense Robert McNamara and had become convinced of the efficacy of McNamara's empirically-based systems-analysis approach to management. This approach was in sharp contrast to the intuitive approach that had traditionally characterized decision making within the Army General Staff. Returning to the Pentagon in 1969 as the Assistant Vice Chief of Staff of the Army, Lieutenant General DePuy instituted approaches to issues that reflected this influence. General DePuy also had a remarkable ability to select and surround himself with outstanding, innovative, and like-minded officers, especially at the Colonel and Lieutenant Colonel level. These included such eventual luminaries as John Woodmansee, Max Thurman, Robert Montague, Colin Powell, and Louis Monterey, all of whom worked for him during his Pentagon years[34] (Kitfield, 1997). All of these experiences and characteristics set the stage for General DePuy's ultimate influence on the Army.

In 1972, as the Army Assistant Vice Chief of Staff, Lieutenant General DePuy was given the task of reorganizing the unwieldy umbrella Continental Army Command (CONARC), which was the controlling organization for all Army units, including schools and training centers, in the contiguous United States. Acting quickly, Lieutenant General DePuy split the command

[34] Thirty of these officers who were directly mentored by General DePuy would become General Officers and seven of them would reach four-star rank. This is an unsurpassed record of Army mentorship.

into two groups – a Forces Command (FORSCOM) and a Training and Doctrine Command (TRADOC). In early 1973, Lieutenant General DePuy was named as the first commander of TRADOC and given his fourth star as a full General.

General DePuy was convinced that the Army's traditional approach to training was broken.[35] Indeed, the experiences of Viet Nam had many on General DePuy's staff concerned that training was a significant weakness within the existing Army. Moreover, the observations and lessons learned from the 1967 Israeli Six Day War and the 20 day Yom Kippur War of 1973 had a profound effect on the thinking of leaders about the Army's readiness and its reaction time. In essence, the nascent TRADOC staff concluded that the Army did not have a viable, unifying training doctrine, and that they were essentially starting over. In so doing, they were not reluctant to look outward – to industry, academia, think-tanks, and any other sources that held promise of a breakthrough to the Army's training issues (Army Training and Doctrine Command, 2003).

General DePuy's selection for his Deputy Chief of Staff for Training at TRADOC was Major General Paul F. Gorman, a brilliant thinker and innovator. Two years earlier, in 1971, newly-starred Brigadier General Gorman was assigned to the Infantry School at Fort Benning, Georgia. Brigadier General

[35] In the early 1970s, the Army was still following the Army Training Plan (ATP), which had been enacted in the 1920s and had changed little since then. The ATP dictated subjects to be taught and the number of hours of instruction for each Soldier. It did not include standards of performance, or even if any demonstration of performance was required – it was entirely built on blocks of instruction to be delivered. The ATP was based on a mobilization assumption, in which large numbers of conscripts would have to be turned into Soldiers. This factor alone gave impetus to the need for change; with the Volunteer Army, the National Defense strategy could no longer rely on mass mobilization as a factor in preparation for warfare.

Gorman had been a brigade commander with the 101st Division in Viet Nam, and felt that the training soldiers received failed to prepare them for that mission. These Viet Nam experiences led him to be instrumental in establishing the *Combat Arms Training Board (CATB)* at Fort Benning.[36] The CATB mission was to identify innovative approaches to training and to stimulate the development of techniques, practices, and devices associated with improving training. Staffed by a small number of mostly young, combat-experienced officers, and civilians deliberately selected from a variety of non-Army related disciplines, CATB's charter was to give impetus and structure to new training ideas. Organizationally, it operated free of the normal hierarchal chain that oversaw other combat developments. Its purpose was to formulate and innovate, but not necessarily to implement (Chapman, 1991).

Retrospectively, the programs that were introduced from June 1973 to July 1977 have become known as "*The DePuy-Gorman Initiatives*" (Chapman, 1991). These innovations significantly, and in many cases, permanently, changed the way the U.S. Army operated. Among the many programs credited to DePuy-Gorman were one station unit training (OSUT), Instructional Systems Development (ISD), performance-based training and testing, task-based training, criterion-referenced instruction

[36] CATB was originally established at Fort Benning as the *Board for Dynamic Training*. It became CATB after becoming a TRADOC organization in July 1973. Under TRADOC, CATB's initial Commander (1973-1975) was Colonel Frank A. Hart, who continued on to have an illustrious and influential succession of assignments within TRADOC and as the Commander of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) in the early 1980s. In October 1977, CATB was merged with the Logistics Training Board from Fort Lee, Virginia, and was moved to Fort Eustis, Virginia, where it became the *Army Training Board*. At that time, it lost some of its original focus. The Army Training Board was inactivated in 1989.

(CRI), Training Extension Courses (TEC), and Soldier's Manuals (SM). These innovations together laid the groundwork for the program that would radically change the Army's personnel evaluation system – the Skill Qualification Test (SQT) program. However, personnel evaluation and job testing were not new to the Army. SQT would replace an existing system – The Army Enlisted Evaluation System (EES).

The Army Enlisted Evaluation System (EES)

In 1957, Secretary of Defense Charles E. Wilson appointed a committee to study military compensation.[37] Chief among the board's conclusions was a recommendation for revision of military compensation from the existing program based on longevity, to one that rewarded performance, recognized advanced skills, and was based on identifying and rewarding high quality personnel. The board also recommended the services implement a manpower management plan to provide a means for effective administration of the new pay plan (Department of Defense, 1957). This gave impetus to several Army initiatives associated with a new personnel management system. Chief among them was the Army *Enlisted Evaluation System (EES)* which included testing soldiers on their jobs. A feature of the EES was that, based on performance on the test, a Soldier could qualify for *proficiency pay* – an additional monthly stipend paid to individuals who performed better than their peers.[38]

[37] This committee was known as the *Cordiner Board* after its Chairman, Ralph J. Cordiner, who, at the time, was the President of General Electric Corporation. This was the first comprehensive review of military compensation since shortly after World War I.

[38] The initial proficiency pay rate was $30.00 per month. In fiscal year 1959, a Staff Sergeant E6 with over 10 years service had a monthly base pay of

The Army implemented EES in early 1959. Although the system changed and evolved during its almost 18-year lifespan from 1959 through 1976, essentially EES was a two-part, composite evaluation program. It consisted of a supervisor's Enlisted Efficiency Report (EER) rating and an annual military occupational specialty (MOS) test designed to measure job proficiency. The EES was the basis for enlisted promotion, job and Army retention, and the awarding of proficiency pay (Department of the Army, 1964).

The MOS tests which were part of EES were developed under the control of the U.S. Army Enlisted Evaluation Center (EEC) at the Army Personnel Command (PERSCOM) at Fort Benjamin Harrison, Indiana. Test items for an individual MOS were developed at the service school responsible for that MOS and sent to the EEC. Almost universally, the item writers were service school instructors, usually non-commissioned officers (NCO).[39] Each MOS test normally consisted of 200 items, of which 125 would be ultimately used as for-record test items (the remainder were pilot items or other non-count items). All tests were paper and pencil and all items were multiple-choice.[40] Items were normally grouped into *Subject Matter Areas (SMA)*, which all addressed like content, such as a piece of equipment (Fabian, 1974).

$255.00, so qualifying for proficiency pay was the equivalent of almost a 12% pay increase (Defense Finance and Accounting Service, 1958).

[39] By the early 1970s, there were over 40 "item writing agencies" (IWA) submitting tests to EEC.

[40] There were also a small number of performance tests developed. There was a typing test for clerk-typists, a taped Morse Code read test for some signal Soldiers, a recording test for court stenographers, and a musical performance test for bandspersons. However, these were never widespread and, even within these MOS, it appears that the performance tests were primarily for demonstration purposes.

MOS tests were administered annually, according to a quarterly MOS schedule. All soldiers who were serving a commitment of two or more years active duty were required to test after they had completed 16 months of active duty.[41] Soldiers tested each year in their primary MOS (PMOS) and once every two years in their secondary MOS (SMOS). Tests were disseminated, controlled, secured, and administered through a network of Test Control Officers (TCO); the TCO were part of the Adjutant General (AG), under the Army G1 function. TCO coordinated directly with EEC for test requisitions and returns. They normally functioned on an installation or geographical area basis, not on a unit level. The TCO would distribute MOS Evaluation Study Guides 90 days before an MOS test date. These would identify, in general terms, the references that were used for item development (Department of the Army, 1964).

Although there was considerable overlap in content and test items, different tests were administered to different pay grades within an MOS. A soldier's score was the total number of items correct, and scores were normed on all soldiers in the same MOS and pay grade who took the same test. In computing a soldier's total evaluation score, the MOS test score was combined with the soldier's numerical EER score. The raw composite score was then converted to an evaluation score on a range of 40 to 160. A score of 100 was the mean of the tested group (MOS and pay grade). A score of 70 was considered as qualifying in one's MOS and a score

[41] Eligibility requirements for testing changed periodically. Normally, only Soldiers in pay grades E3 through E8 were tested. E9 were exempt from testing after an initial qualifying test. From 1964, Soldiers in Reserve status (Army National Guard and Army Reserve), E4 and above, were also tested, although not on an annual basis.

of 100 was generally considered as necessary for promotion.[42] Proficiency pay cutoffs were usually established separately for combat arms and non-combat arms MOSs; however, the scales and rates fluctuated over the years. Normally, proficiency pay was only awarded to those in the top one-third scores in any group (Bryne, 1971).

The MOS test was a significant system, and the fact that it endured through 18 years indicates its important role in the post-World War II Army. Ultimately, however, the program was brought down by inherent problems. Although attempts were made throughout its existence to enhance its relevance to training and unit readiness, the MOS test was primarily a personnel tool to support the promotion and retention system and its impact on training and readiness was almost insignificant. A further reason for the decline in support for the MOS test was that, primarily due to inflation in the EER, the EES was no longer useful in discriminating among soldiers for promotion (Bell & Cocke, 1977a). Additionally, because the content of the MOS tests came primarily from the schoolhouses, there were widespread complaints that the tests were not valid measures of what field soldiers did on the job. Coupled with this was the complaint that multiple-choice tests—what a soldier *knew*—were not an adequate measure of what a soldier could *do*. Finally, however, the most significant factor in the decline of the MOS test was that it could not logistically be administered to soldiers in Viet Nam. Therefore, from 1965 on, large segments of the Army did not take MOS tests or only took them sporadically. By the early-1970s, the large scale Viet Nam manning levels had passed, but by then the Army's commitment

[42] Minimal hurdles were established in both the normative MOS score and the weighted EER score that had to be met before the composite score could be computed.

to, and involvement in, the MOS test had waned. The Army was ready for a change to its personnel job evaluation system.

The SQT

The SQT program lasted from 1974 until the ultimate demise of the derivative Self Development Test (SDT) in 1994. However, during this 20 year period, the SQT was never a single system, and almost from the start, it was in a constant state of flux. Although legions of soldiers and others participated in SQT first-hand during its lifetime, their experiences and recollections differ markedly, based on when they served. Moreover, the changes in the SQT program were not always distinctly delineated and documentation was often not precise. Nonetheless, in retrospect, it is possible to identify three periods of the SQT era:

- The Initial SQT Period: 1974 -1979
- The SQT Retrenchment Period: 1980-1983
- The Post-SQT Period: 1984 – 1994[43]

The Initial SQT Period: 1974-1979

The precise point of origin of SQT will probably never be known. The earliest formal record seems to be a concept paper titled *Proposal to Implement Performance Testing for Enlisted MOS* that circulated as a draft in TRADOC in 1973 and eventually was approved at Department of the Army (DA) level in 1974 (Campbell, 1994). As noted earlier, the concept of

[43] Unless otherwise specifically cited, the information in the following sections is based heavily on *The Army Skill Qualification Test (SQT) Program: A Synopsis*, (Campbell, 1994). This review and recording of SQT events was written by Roy C. Campbell with the Human Resources Research Organization (HumRRO) in 1993. Mr. Campbell was involved in SQT development starting in 1973 and throughout the first 12 years of the program.

performance testing existed even in the MOS testing under EES and was an integral part of the emerging performance training initiatives. Although the concept paper had no identifiable authors, it likely emerged from someplace in CATB.

During 1974 and 1975, the major conceptual components of the SQT took shape, many of them directly adapted from the corresponding training changes that were then occurring. The SQT concept was organized around the following specifics (Maier & Hirshfeld, 1978):

- Performance testing would be the primary mode for the SQT. Even "written" tests would be structured in a performance context, in which the soldier would have to do something, rather than simply know something.
- SQTs would be criterion referenced. The criterion would be the Army job domain, and soldiers would be measured based on the performance criterion, not on how they did compared with other soldiers.[44]
- SQTs would be job relevant. Critical task lists would be developed for each MOS and skill level and each task would be defined by the behaviors necessary for task performance. Tasks would be described in job-specific Soldier's Manuals and all tests would be based on the Soldier's Manual. Testing would drive soldier and commander attention and concentration.
- Tests would be validated prior to administration following specific validation procedures. (Test validation had been a contentious issue in the MOS test.)

[44] While the tests were criterion based, the Army never adopted a criterion-based personnel system. Normative rankings would still be used for such decisions as promotions, MOS and Army retention, proficiency pay and other personnel issues.

- The SQT would be the cornerstone of a training management and evaluation program.

SQT Organizational Changes

In September 1975, a Memorandum of Understanding (MOU) between the Commander, Military Personnel Center (MILPERCEN) and the Deputy Chief of Staff, TRADOC was executed that allowed for the transfer of testing functions to TRADOC. In return, TRADOC agreed to provide MILPERCEN with reliable and valid evaluations that could be used for the purposes of personnel management.[45]

To support SQT conceptual and implementation development, and to coordinate SQT requirements within the service schools, TRADOC established several organizations. The primary organization was the *Training Management Institute* (TMI), which was established at Fort Eustis, Virginia in 1975 to act as a developmental and quality control agency throughout TRADOC. Although its role in TRADOC training eventually expanded considerably, its initial focus was on SQT. From its inception in 1975 until April 1978, it was commanded by Colonel Robert P. Dirmeyer,[46] who personally spearheaded most of the early SQT conceptualization, development, and implementation. There was also a subordinate organization within TMI—the *Individual Training and Evaluation Group* (ITEG)—which was directly focused on SQT development and

[45] This was significant. For the first time, personnel testing would be part of the Army's training sphere, not a personnel function and not under control of the Army Personnel Command.

[46] Colonel Dirmeyer had been an artillery commander in Viet Nam in the 101st Division where he worked closely with then-Colonel Paul Gorman. He later rejoined with then-Brigadier General Gorman at the CATB at Fort Benning, Georgia. Colonel Dirmeyer was personally selected by Major General Gorman and by General DePuy to head the SQT program, starting in late 1973, at Fort Eustis, Virginia.

implementation. Staffed with both professional civilians and selected military, ITEG performed a wide range of functions including SQT conceptual development, test developer training, test production, and SQT analysis[47] (Chapman, 1991).

The requirements of SQT were totally new, and the procedures for developing and validating the tests were unprecedented. To support the field development agencies, Colonel Dirmeyer's ITEG unit and elements from TMI developed a three-week workshop which was a criterion referenced, self-paced, modular training program, covering all aspects of SQT development and validation. During 1976, this workshop was delivered first as part of a train-the-trainer program and subsequently to over 800 participants at 20 different TRADOC SQT development locations (Campbell, Ford & Campbell, 1978).

As has been noted previously, the SQT was not a stand-alone program. A significant restructuring of the Army's Enlisted Personnel Management System (EPMS) started in 1974 with a 3-year implementation plan. Its objective was to integrate subsystems for training, MOS classification, evaluation, and promotion, and to standardize career patterns. The EPMS provided for the grouping of MOS into Career Management Fields (CMF) for control and administrative purposes. In 1975, there were 451 MOS in 36 CMF[48] (Cocke, 2000). Each CMF and MOS was assigned to a proponent—normally a service

[47] TMI was redesignated as the *Training Developments Institute* (TDI) in May 1977. In 1984, TDI was redesignated the *Training Technology Agency* and its mission changed. ITEG was redesignated as a Directorate (ITED) in 1977.

[48] During Viet Nam, the number of MOS was around 490. Post-Viet Nam, efforts were made to consolidate and eliminate MOS and the numbers incrementally declined over the years. By the time of the initial implementation of the SQT in 1977, there were somewhat fewer than 450 MOS. By way of comparison, in 2008 there were approximately 170 MOS in 29 CMF.

school—for doctrine, training, documentation, and, eventually, test development. In 1975, there were 20 proponents, of which 18 were TRADOC institutions and two—the Academy of Health Sciences and the School of Music—were non-TRADOC. The EPMS also emphasized the grouping of pay grades into skill levels (SL) for career management and the skill level concept became critical to the SQT (See Table 9.1).

Table 9.1. Pay Grades and Skill Level Equivalents

Pay Grades	Skill Level
E1,E2,E3,E4	SL1
E5	SL2
E6	SL3
E7	SL4
E8,E9	SL5

Each skill level in each MOS had its own list of individual tasks that comprised the domain of individual skills that the incumbent soldier must be able to perform.[49] The definition of a task was rigid and derived from the ISD as "*...a highly specific action that has a definite beginning and end, is performed in a relatively short period of time, is observable, measurable, and independent of other actions*" (Branson, 1975, p. A-13). Any activity that did not meet this definition could not be included in the individual task list. The tasks, their performance steps, conditions, and standards were detailed in the Soldier's Manual for each MOS and skill level. Production of Soldier's Manuals was a monumental effort of the proponents in the early 1970s, but both the Soldier's Manuals and the supporting task analysis were essential elements of the SQT program.

SQT Design

As the initial SQT concept evolved, there were several features that were incorporated into the design. These were:

[49] Under the doctrine of task assignment and responsibility, task accountability accumulated as Soldiers moved up in Skill Level. For example, a Soldier at SL3 was still responsible for all the SL1 and SL2 tasks in that MOS; no tasks were ever "left behind."

- A primary purpose of the SQT was to determine readiness for promotion. A soldier would have to "verify" at their current skill level and "qualify" for the next higher skill level.
- SQT would be distinct by MOS and skill level, and, within an MOS and skill level, by duty position or equipment.
- The focus would be on measuring performance. The design of the SQT would allow for multiple modes (components) to capture performance. (Three components were designated – a written test, a hands-on performance test, and a special performance component; these are described in detail in subsequent sections.) The test would provide for equivalency between components while allowing some flexibility on which components were used.
- The SQT was to be task based. All scoring would be on overall task proficiency, not points or number of correct responses. Tasks would be organized into "scorable units." Task proficiency would be based on criterion, scored as "GO" or "NO-GO".
- The SQT would be an "open" test. Soldiers would be provided an SQT Notice 90 days in advance of testing which would tell them exactly how they would be tested and what their test and performance requirements would be.

As will be described, implementing this design led to a complex, sometimes confusing, structure for test development and scoring. The complicated requirements sometimes were ultimately too burdensome on many developers who were

charged with implementing a strategy that they did not fully comprehend. While the design was well intentioned and supportable as a test program, it lacked the simplicity needed for large scale implementation.

Initially, each skill level for each MOS was to have its own distinct SQT. SQTs were also designated numerically; however, as shown in Table 9.2, the skill level and SQT numbers did not match. This was because there was no SQT1 for any MOS. This somewhat confusing numbering system came about because of the requirement that soldiers would qualify for the next higher skill level through their SQT performance. A SL1 soldier would qualify for SL2 by taking an SQT2.

Table 9.2. Pay Grade, Skill Level, and SQT Numbering

Pay Grade	Skill Level	SQT Number
E4	1	2
E5	2	3
E6	3	4
E7,E8,E9	4,5	5

To understand how the "verify/qualify" requirement was built into the test, it is necessary to detail the structure of the initial SQT. The basic building block was the *scorable unit* (SU). Normally (there were some exceptions), an SU equated to an individual task. Each SQT was to be made up of multiple SU which were spread across the three components of the SQT. The

minimum number of SU required for an SQT was 40 and the maximum number that could be included in an SQT was 86.[50]

To build in the "qualifying for next higher skill level" feature, the requirement was that the SQT be constructed with two-thirds to three-quarters of the SU being tasks at the current skill level, while the remainder must be from the next higher skill level. For example, an SQT2 with 86 SU could be based on 57 SL1 tasks and 29 SL2 tasks (Individual Training and Evaluation Directorate, 1977).

This concept never worked as planned. Because proponents began with development of SQT2 and worked up, the SQT5s were never fielded, so a distinct test for E7, E8, and E9 never became a reality. In fact, few MOS ever fielded an SQT4; the emphasis was always on producing SQT2 and SQT3. Moreover, the structure of most MOS (the same in 1977 as it is today) was that the preponderance of tasks in an MOS were located at SL1; many MOS, including dense MOS such as tank crewman, motor transport operator, and wheeled vehicle mechanics, had between zero and five SL2 tasks. Although the "next higher skill level" inclusion was a strong selling point of the SQT in the early 1970s, it was never a significant reality of most of the SQTs. At some point in 1979, the SQT were renumbered to designate SQT1 at SL1. However, in many MOS, only a single SQT was actually produced and all soldiers who were tested, regardless of rank or skill level took the same test.

There was also a provision in the SQT for *tracking*. Tests could be tracked when there were job, task, or performance variations

[50] The SQT was, potentially, a very large and very comprehensive test. Although tests did not reach the maximum number of SU, the idea of annually testing the equivalent of 86 individual tasks was unprecedented. In many cases, this would exceed the entire inventory of a Soldier's job tasks as defined by the Soldier's Manual.

dictated by different duty positions or equipment within the MOS. The usual approach was to provide a shared core of SUs that applied across the MOS and then to split into tracks for variations within the MOS. In practice, the tracking by duty positions was rare. However, the tracking by major items of equipment was common in many MOS and was widely applied. For example, a field artillery SQT could be tracked to include items on each of the three major artillery systems in use at the time.

The SQT Components – HOC, WC, PCC

As noted, there were three components to the SQT – the hands-on component (HOC), the written component (WC), and the performance certification component (PCC). Each of these was designed to offer a particular contribution to the SQT.

The HOC. The most visible component of the SQT was the HOC. This was a performance test, usually administered under field conditions according to strict rules and under precise conditions (including scripted instructions) and scored according to exact standards. The focus during development was on the scorer, who would be an NCO with little training as an SQT scorer, and on ways of enhancing scorer reliability.[51] The HOC could include up to 16 individual tasks (SU). Each SU was comprised of up to 20 Performance Measures (PM) which were individually scored. The PMs were either individual steps (procedures) or outcomes (products) required in task performance.[52] Each PM was scored GO or NO-GO and,

[51] Most NCOs, by their experiences and inclinations, consider themselves to be *trainers* foremost. To turn them into objective observers and scorers was challenging.

[52] The 20 PM limit was because the results of a Soldier's test were to be later transcribed to a machine scannable scoresheet and became part of a permanent record. The machine scannable scoresheet could only

generally, it was required that a Soldier receive all GOs to receive an overall GO on the SU. The HOC had to be administrable within four hours to an individual soldier and within five days to all the soldiers in a given MOS in a given unit (usually a battalion). There was no requirement that each SQT have an HOC. However, because of its emphasis and visibility, almost all early SQTs had some hands-on tests (Individual Training and Evaluation Directorate, 1977).

The HOC was set up to be group administered within a unit. Testing stations were established and replicated as necessary to handle the projected soldier load. Soldiers rotated through the stations and testing was usually an all-day affair. For larger MOS, testing could consume several days. All this was spelled out in a *Manual for Administration: HOC* which was prepared by the proponent. This covered all detailed equipment, personnel requirements, and scorer training procedures. Because the HOC was often equipment dependent, the related equipment requirements for testing were often extensive.

The WC. The WC was a multiple-choice based test, but was very unlike the old MOS multiple-choice tests under EES. The SQT WC test was built on SUs that were based on tasks, normally one SU to a task (some complex tasks could have two or three SUs). Each SU could contain up to 10 items (questions) and each item could have between 2 and 10 responses (alternatives).[53] Items could require more than one response (e.g., "select all correct"). Unlike the HOC (which generally

accommodate 20 entries per SU. To accommodate particularly long or complex tasks, the manual scoresheets could be built with a series of subset scoring measures under the primary PM. However, only 20 primary PM could be permanently preserved.

[53] As with the HOC, this structure was based on the design and layout of the machine scoresheet on which Soldiers recorded their responses.

required all GOs to pass), part of the development requirement for the WC was to establish passing standards for each SU. Guidelines provided were that if there were three or fewer items in the SU, the Soldier had to get them all correct to pass the SU. Above three items, a scale was suggested that ended up with seven correct at the 10 item level, although developers were free to establish different passing standards. Similarly, passing standards were established for scoring the alternatives within the individual items. This generally applied where multiple responses were indicated because the scoring involved all the alternatives – both marked and unmarked. For example, in an item with 10 alternatives, four of which were correct (to be marked), the guideline standard was eight, in a combination of correctly marked and unmarked responses (Individual Training and Evaluation Directorate, 1977).

The individual items that made up the SU in the WC were to be "performance" or "performance-based." That is, they either required the Soldier to "do" something (such as compute a back azimuth) or to answer an item about a step in a procedure (such as how much insulation to strip when connecting field wire). Guidance proscribed general knowledge items such as nomenclature or items not directly related to a performance step in a task. Every SQT was required to have a WC. The WC could take up to four hours to administer to an individual soldier.[54]

The PCC. The PCC was designed to test those tasks that would normally be tested in the HOC, but could not because of

[54] The WC was also group administered. Some of the WC had a distinct performance aspect. For example, an early Air Defense WC test required Soldiers to identify aircraft from slide-photos that were projected to the examinees for 4 seconds each. Soldiers marked their choice for each photo from a list of 10 identifiers.

equipment, conditions, or time constraints. Generally, the construct of a PCC was very much like the HOC hands-on test, with performance measures, specified conditions, instructions to the examinee, and instructions to the scorer. However, it was meant to be administered by the soldier's supervisor at any time during the year. The intent was that a PCC score could be recorded when the occasion for task performance occurred on the job, and often under less stringent conditions than the HOC tests. A PCC could contain from zero to 10 SU.

The PCC was badly served from the start. In fact, it never got much of a try-out and developers were actively discouraged from using it: "*A task should be assigned to the PCC only as a last resort*" (ITED, 1977, p. 3-25). Primarily this was because it was felt that the PCC would lack standardization and the scoring would be unreliable. Although some proponents later included a few job-specific performances in the PCC, the primary application was to include already standardized events—specifically, Weapons Qualification and the Army Physical Fitness Test (APFT). A soldier's last performance on these annual requirements would be translated to an SU credit: a common example for Weapons Qualification would be: Expert = 3 SU, Sharpshooter = 2 SU, Marksman = 1 SU, Unqualified = 0 SU. The "score" would be recorded on the PCC score sheet as total GOs.

Scoring the SQT

In scoring the SQT, each passed SU was awarded one point. All individual scores were converted into percentages. Soldiers were assigned to one of three levels based on their performance:

- 0%-59% - Unqualified
- 60%-79% - Verified at current skill level

- 80%-100% - Qualified for promotion to next higher skill level

Soldiers who failed to make 60% had to be retested again the next year. A soldier who had two successive Unqualifieds was required to meet with an MOS reclassification board or was barred from reenlistment. Soldiers who Verified did not have to test again for two years; however they could retest voluntarily to try to raise their score. Soldiers who Qualified could keep their score for two years or until promoted. It should be noted that being Qualified did not mean an automatic promotion: Soldiers still had to meet time in-service and time in-grade requirements, have a commander's recommendation, pass promotion boards, meet promotion point cutoffs, and/or have vacancies within their MOS. Nor was an SQT Verify an automatic block to promotion: These Verified Soldiers fell into a secondary promotion pool (not to be confused with a Secondary Promotion Zone), and their SQT requirements could be waived. However, only one career SQT waiver was allowed. Finally, soldiers who, though no fault of their own did not have an SQT within two years, or whose MOS did not have an SQT, were not automatically penalized. Such soldiers received a "credit" of SQT category points based on a scale of points received in other promotion consideration categories (Department of Army, 1978).

To be eligible for an SQT, a soldier had to have one year time in-service and have held their primary MOS for at least 90 days.[55] By the start of FY1977, the Army enlisted strength was

[55] There was also a requirement that Soldiers take the SQT in their secondary MOS (SMOS), which was awarded upon attaining E6 promotion list status. Under the MOS test, Soldiers were required to test in their SMOS every other year. Under SQT, Soldiers were required to test, and Verify, in their SMOS once during their career.

at about 680,000 and about 450,000 soldiers were in pay grades E3 through E6 – the prime target for SQT (Department of the Army, 1977). The goal was to have all 36 CMF functional under the EPMS (including established task lists and published Soldier's Manuals) by October 1977.[56] The old MOS test would not be administered after 1 January 1977, and SQT was to start with the first MOS in April 1977, with five more CMF to have fielded SQT by the end of 1978. Total SQT fielding for all CMF was targeted for 1980.

An SQT (from MOS 11B - Infantryman) from CMF 11 (Maneuver Combat Arms) was administered on schedule in April 1977, followed by tests in CMF 16 (Air Defense) and CMF 95 (Law Enforcement) in September 1977. By late 1977, proponent agencies for most CMF were working on some phase of SQT development and, in some cases, implementation. But by 1978 serious problems were emerging with SQT and the overall schedule for the program slipped rapidly and drastically.

The SQT Retrenchment Period: 1980-1983

By 1980, the SQT program was in trouble. There were three major levels of problems that, unfortunately, permeated the entire system. To comprehend these problems, it is necessary to understand some of the complexities of how the program was to function. As indicated previously, the overall responsible organization for SQT was the Individual Training Evaluation Directorate (ITED) at Fort Eustis, Virginia. SQT products were developed by the TRADOC proponents (usually the service schools), and reviewed, approved, and distributed by ITED on a quarterly test schedule to approximately 60 test control officers (TCO) world-wide. (In the late 1970s, about 65% of the Army

[56] In actuality, about 80% of MOS had their Soldier's Manuals and were under EPMS by 1977. (Department of the Army, 1979)

was stationed outside of the Continental United States.) The TCO coordinated with the units in his or her jurisdiction for distribution of SQT Notices to individual soldiers and for scheduling the tests. Units were responsible to schedule, support, and administer the tests and provide the raw results back to the TCO. The TCO collated the data and sent them back to ITED where they were entered and verified and the test scores were produced. ITED was then to provide results back to the individual soldier and the supporting personnel command for entry into the soldier's records, as well as a roll-up report to the soldier's unit. ITED performed two other important but dissimilar functions. First, as the repository of all the test data, it had to perform analytical and confirmatory analysis. Second, ITED was a distribution center; it had to print, assemble, bind, package, and distribute SQT Notices and test materials on time and in quantities for the TCO to make necessary distribution.

The first level of problems started at the proponents. SQT required not only development, but also piloting and validation. The proponents had to prepare the SQT Notice, and the HOC required a separate *Manual for Administration* covering all aspects of set-up and administration instructions to the field unit under a hypothetical "best case" administration scenario. In order to meet ITED production requirements, all materials produced had to meet stringent and uniform camera-ready mechanical layout standards. New staff skills and expertise were required, particularly in the development of hands-on tests, and, despite increased numbers of personnel and an intensive TMI training program, the impact on the proponents was immense. In retrospect, there was a significant learning curve that had not been anticipated. As was outlined in the preceding design description, the SQT was complex and not always fully understood. Moreover, there was no other experience for preparing and conducting performance testing on such a scale

from which to draw—not academic, industrial, or other DoD. The end result was that proponents were simply unable to produce effective, mistake-free tests in the numbers required by the original time schedule.

The second level of problems existed within ITED. The production and distribution requirements were massive, and field deliveries of SQT products were not always timely or accurate. ITED's oversight, proponent assistance resources, and quality control were overtaxed, allowing some flawed products to get into the system, which resulted in the invalidation of some of the tests.[57] Further, one of the promised features of SQT was that results could be aggregated at the company, battalion, brigade, and division levels, giving commanders a precise picture of training needs. The feature never functioned effectively, which negated one of the training management advantages promised of SQT.

The final, and potentially most critical problems, occurred in the field. As often happens, SQT requirements were superimposed on units without any lessening of already burdensome demands and missions. Units not only had to provide soldiers to be tested, they also had to provide equipment, scorers, and other administrative and logistical support. A battalion conducting a large scale SQT often had to shut down for 2 to 3 weeks just for SQT administration. As the program became more widespread and repeated several cycles, the situation for the units got worse

[57] Between 1977 and 1981, 6% of the SQTs that were fielded were subsequently invalidated. Most of these were WC tests. Ultimately, most problems were doctrinal and were the proponent responsibility. Some problems were endemic – five consecutive WC SQT fielded for the supply specialist MOS were invalidated because they were based on obsolete doctrine or procedures (United States General Accounting Office, 1982).

instead of better.[58] The pressure grew for units to "do well" on SQT, which increased the requirement to allocate more time to SQT preparation. Commanders' perceived that they were being judged on SQT outcomes and that the threshold for a successful SQT was very high. Feedback from the field, both official and back-channel, was that field units could not sustain the burden that SQT imposed.

As it became apparent that the 1980 target date for complete Army-wide SQT implementation could not be met, the deadline was delayed until 1981. When this date also could not be met, it was advanced again, until the Army finally stopped announcing target dates. In fact, about 15% of the MOSs never got an SQT (Oland, 1995). These delays caused another problem—the perception that SQT was "unfair" because some MOS had SQT and some did not. This issue was frequently publicly cited, even at very high Army levels, throughout the SQT period. In reality, at least from a promotion standpoint, no group of soldiers was at a disadvantage because they had, or did not have, an SQT. (Promotion competition was for vacancies within an MOS, not between different MOS). Nonetheless, the perception was the reality, accurate or not.

Although problems existed at all levels, it should not be inferred that the SQT system was in total chaos. The majority of the tests were very good and large groups of soldiers were being evaluated effectively. Most importantly, deficiencies and proficiencies in individual performance and training were being identified. The consensus, even among critics, was that the program was, in concept, a good one. But the problems with

[58] Each MOS was allocated a three month window in which the SQT was to be administered. As most units had a variety of MOS, units were eventually faced with some level of testing year-round.

SQT had to be addressed, so revisions were made to how SQTs were organized, developed, and administered. Many of the changes had to do with labels and nomenclature—not a major issue, but one that makes subsequent tracking of the SQT history confusing. Most of these initial changes occurred in 1980; a few were incrementally applied in following years. The following list summarizes those changes (Department of the Army, 1980):

- The requirement to include tasks from the next higher skill level in a test was dropped. The test for a skill level would cover tasks designated for that skill level. Tasks from lower skill levels could be included in a test designated for a higher skill level, but higher skill level tasks were not to be included in lower skill level SQT. (For example, a test for E6 (SL3) could have tasks from SL1 and SL2. However, a test for E5 (SL2) could not include E6 tasks from SL3.)
- The numerical SQT designation was dropped. An SQT would be called by the skill level for which it was intended (e.g., SQT SL1).
- The Verify (at 60%) and Qualify (at 80%) distinction was dropped. SQT results would be Pass-Fail with the Pass point set at 60%. Initially, there was an effort to award a set number of promotion points at passing (60%) and to double those points if a Soldier achieved a 90% score. However, there was no scale in-between and this policy was short-lived. Ultimately, and for the remainder of the program, the percentage score on the SQT was doubled and converted to promotion points on the Promotion Point Worksheet (PPW). The range for SQT was from 120 to 200 points out of a total PPW

possible of 1000 points. The Soldier who got less than 60% received zero SQT promotion points.

- The concept of the Scorable Unit (SU), which had been used to establish equity between the different SQT components, was dropped, as was the term itself. Although the task-based concept of SQT continued, scoring and credits would be on a points basis rather than on specific tasks.

- The hands-on test (HOC) remained basically unchanged in both format and content. However, new guidance was issued specifying that a typical unit must be able to test all its soldiers within an 8 hour period. This had the effect of reducing the hands-on test to one-half to one-third of its previous scope. Proponents were also tasked with producing an *Alternate Hands-On Component* (AHOC) test. The AHOC was essentially a written test that covered the same tasks that were in the HOC. Under certain guidelines, units could elect to administer the AHOC in lieu of hands-on testing.

- The term Written Component (WC) was dropped and replaced by the term *Skill Component* (SC). Although the SC was essentially a written test, the stated intent was to make the test more visual, with pictures, graphics, and diagrams, and less of a reading requirement. The SC was also standardized as a four-alternative, single-correct, multiple-choice test. The size of the SC was to be much smaller than the original Written Component (which had a 4 hour time limit). Guidance was that the SC had to be administrable within one hour for combat arms MOSs and within two hours for the technical and support MOSs.

- The term Performance Certification Component (PCC) was dropped and replaced by the term *Job Site Component* (JSC). In principle, the JSC was to operate like the intent of the PCC, but there were significant differences. The JSC tasks were selected by the proponent and published in a JSC Booklet to be provided to the Soldier's immediate supervisor. The JSC Booklet retained a hands-on flavor in that it listed task performance requirements as specific performance measures. However, there was less emphasis on the conditions and test-like requirements of administration. The supervisor could evaluate the task any time the Soldier performed it and there was no limit on the number or timing of the observations. The Soldier did not even have to be aware of the evaluation. There was also a provision that if a JSC task could not be administered on the job, the Soldier could get be credited SQT points for completing a Training Extension Course (TEC) lesson or an Army Correspondence Course (ACC) lesson that included the task.

- The different components of the SQT were to be weighted. The SC (written) was to be weighted less than the HOC and JSC. Other than this, exact weightings were up to the proponent.

The Post-SQT Period: 1984-1994

The revisions instituted in the SQT during the early 1980s bought the program some time, but did not erase the problems that caused the changes to be made in the first place. If anything, as the SQT spread to more and more MOS, the problems began to magnify. Proponents found that, even with second and third generation tests, the manpower demands for development and try-out were extensive. Many proponents were

complaining that the cost of SQT development was unsupportable. In some cases, proponent in-house test development expertise was still lacking, resulting in continued field complaints about the quality of the tests. Some proponents failed to adequately analyze job requirements and test relevancy again became an issue, adding to user dissatisfaction. As the tests proliferated to more MOS, ITED's resources were spread even thinner, further affecting the products fielded. But the greatest impact continued to be on the field units and the SQT program was eliciting strong, vociferous reactions from some commanders. Even with the reduction in the scope of the tests, the overall perceived effect was that there was more testing going on. The major commanders saw greater and greater demands on resources and SQT figured significantly in these perceptions. Both soldiers and leaders still feared (however erroneously) that the unevenness in how the program was implemented would disadvantage large numbers of soldiers. Finally, and significantly, leaders were not finding any advantages in supporting the program, in terms of evidence of readiness or training needs.

The problems were elevated to a higher profile with the release of a Government Accounting Office (GAO) study in 1982 that was highly critical of the SQT program (GAO, 1982).[59] Based on the initial draft of the GAO report, Congress, in December 1981, halved funding for SQT from $18 million to $9 million. Although they later restored an additional $4 million, this was a serious call to the Army to pay attention to the SQT

[59] The GAO report covered the SQT prior to many of the changes covered in the preceding section having been implemented. However, the GAO issued a subsequent report in August 1982 that was also critical of the Army's response to its initial report and maintained that no significant changes, particularly in the area of cost reductions, had actually been made (GAO, 1982, August 18).

shortcomings (Department of the Army, 1988). While some of the rationales and assumptions of the GAO study were disputable or even refutable, they raised legitimate issues, including the following (GAO, 1982):

- Test results did not accurately indicate a soldier's ability to perform critical job tasks because only a limited number of tasks were tested.60
- SQT were being used as once-a-year events, instead of being the culmination of a year-round training program.
- Promotion decisions based on SQT results created inequities among soldiers.
- Test results were not used routinely to measure either individual proficiency or training needs at the unit level.
- SQT was costly. Although costs were difficult to pinpoint, the GAO estimate was around $25 million annually

Although the GAO report is often cited as the "turning point" of SQT—and its Congressional reaction insured a strong Army response— most of the continuing criticism of the SQT program and its burden was, in fact, being generated from within the Army. Dissatisfaction was widespread, including from the proponents, but most noticeably and persistently were the ongoing complaints from field commanders. Moreover, there had been widespread turnover in the Army leadership since the inception of SQT and, with command changes, policy

[60] Somewhat paradoxically, the GAO criticism was that the SQT did not test enough of the Soldiers' job. Elsewhere in the report they echoed the Army commanders' complaint that SQT administration was taking too much time and resources.

changes inevitably follow.[61] Starting in 1983, there was a major redirection of the Army's approach to soldier testing.

In 1984, the Army officially dropped the full, three-component SQT program that had been in use since 1977. In its place, the Army adopted the *Individual Training and Evaluation Program* (ITEP). Although ITEP still incorporated a test, and still was intended to be training oriented while providing a basis for personnel actions, it was a significant departure from even the modifications to SQT of the early 1980s. The ITEP was part of the *Battalion Training Management System* (BTMS), and, as such, its primary application and output were decentralized. Collection and centralization of individual performance measures were sharply curtailed under ITEP. The following list summarizes the major features of ITEP, as compared with the predecessor SQT program (Department of the Army, 1985).

- Both the Hands-On Component and the Job Site Component were eliminated.
- There was to be a written portion of the ITEP. That written test was now officially called the *Skill Qualification Test (SQT)* and was the only part of Soldier evaluation to have the SQT designation. The SQT was MOS specific and organized around job tasks. The SQT was not to exceed two hours administration

[61] In March 1983, General William R. Richardson was installed as TRADOC's fourth Commanding General. The climate and leadership focus of TRADOC had changed since General DePuy's initial tour ten years previously. General Richardson's priorities were with force modernization and Reserve mobilization. Additionally, TRADOC prestige and influence had declined perceptibly in the early 1980s and much of General Richardson's focus was toward reinvigorating TRADOC's mission and morale. He believed that the heart and future of TRADOC was in new organizations being established at Fort Leavenworth, Kansas. Individual testing was no longer a prime TRADOC priority (TRADOC, 2003).

time and items were to be limited to multiple-choice items with four alternatives and one correct answer. Tasks to be included in the test were to be based on a sampling from the Soldier's Manual. There were generally to be four levels of a test (SL1 through SL4) that were to be administered to all personnel, E1 through E7.[62] Soldiers were to be tested once each fiscal year during a single, 3 month, Army-wide SQT window. A score of 60% was required for passing and scores would be reported and entered on the Solder's individual records. A Soldier's scores would also be reflected in points on the PPW. The SQT was the only ITEP indicator of individual proficiency that was reported or recorded outside of the Solder's unit.

- A *Common Task Test* (CTT)[63] was created, consisting of 7 to 10 SL1 tasks from the Soldier's Manual of Common Tasks (SMCT), to be selected by TRADOC. The CTT was to be a performance test, administered hands-on and based on performance measures published in the SMCT. TRADOC also provided a written version of the test to be administered if conditions precluded hands-on testing. The CTT was to be administered

[62] There could be, and was, considerable overlap among the test items in the four levels of tests.

[63] "Common tasks" were a spin-off that originated as an adjunct of SQT in the early1970s. The original concept was to provide a list of tasks for Soldiers who did not yet have a Soldier's Manual or whose MOS tasks did lend themselves to hands-on testing. The idea was to select tasks that could reasonably be expected to apply to all Soldiers regardless of MOS. The first Common Tasks Manual was published in May 1977 and contained 28 SL1 tasks in first aid, communications, security, NBC, camouflage and concealment, M16 rifle, and physical fitness (Department of the Army, 1977b). By the early 1980s, the Common Task concept had evolved to become a requirement for all Soldiers. Today, it has grown considerably in scope and is one of the universal pillars of all Soldier training.

annually to all soldiers, E1 though E7. The test could be administered any time during the year. The administration, including retesting, was up to the unit. Use of the results, both individually and collectively, was the prerogative of the unit commander; no centralized reporting was done and no record of performance was officially maintained for any Soldier.[64]

- *Commander's Evaluation* (CE). This was perhaps the most nebulous part of ITEP. The CE was to be a performance evaluation of individual Soldier proficiency. The CE provided for the Soldier's supervisory chain to evaluate the Solder's ability to perform on "mission-related tasks" selected from the Soldier's Manual. These tasks were not prescribed and could vary with each unit, supervisor, or Soldier. The evaluation could be formal or informal. It was the responsibility of the supervisor to tell each Soldier what tasks had been selected for his or her particular CE, to conduct the evaluation, and to critique each individual's performance. Instruction on the performance of the task, by the supervisor or others, could be done individually or in a group and could constitute the evaluation. For soldiers in SL1 and SL2 (E1-E5), supervisors were required to maintain Job Books that listed for each individual the tasks in the CE and the date and "results" of the last evaluation. There was, however, no reporting

[64] The CTT, in basically this same structure, has survived to the current date. TRADOC continues to publish an annual CTT requirement list and updates it each year. CTT is a valuable tool for training emphasis. However, CTT is not a "test" under any of the accepted concepts of testing – it lacks standardized administration and performance requirements, uniform scoring, and controls for administration and recording. While it remains an important training tool within the Army, it is a "test" in name only.

requirement and there was no officially mandated use of the individual CE results.

The controversy about SQT did not end with the introduction of ITEP. In 1985, soldiers provided feedback that expressed dissatisfaction with the written SQT. Test scores had declined under ITEP and soldiers were convinced that they would suffer promotion, reenlistment, and retention problems as a result. Ironically, their complaints were the same as those of the pre-SQT period – that performance ability should be considered as well as "text-book" knowledge. TRADOC sought to resolve this discontent with a new proposal for hands-on testing, but the Major Army Commands (MACOMs) resisted strongly and the revision was shelved pending further study and eventually died (Cocke, 1995).

While SQT did not remain completely static through the late 1980s, it essentially remained as a two-hour written MOS test. There was some emphasis on making the test easier and, in fact, SQT scores continued to rise into the 85%-95% levels, at least in some MOS. Many proponents took on SQT as a maintenance activity and many tests were minimally changed from year to year. Nonetheless, there was still dissatisfaction with the SQT. Part of the issue was that SQT was still primarily a unit responsibility and commanders were still held responsible for test preparation. In 1990, in reaction, principally, to continued field feedback, it was proposed to eliminate the SQT and replace it with a self-development system. Sergeant Major of the Army (SMA) Julius Gates was especially critical of the SQT, as were Command Sergeant Majors at TRADOC, FORSCOM, and other major command positions. Army Chief of Staff General Carl Vouno, (a former TRADOC commander who had succeeded General Richardson at TRADOC in 1986)

supported these senior NCOs and made the decision to eliminate all testing for pay grades E1 though E5[65] (Fisher, 2001).

In 1991, the terms *Skill Qualification Test* and *SQT* left the Army lexicon. They were replaced by the term *Self-Development Test* (SDT) in October 1991. While SQT was MOS specific, SDT was touted as a test that incorporated training and leadership and, as a result, was more suited for NCO development. The SDT was a two hour, 100 item test, with 20 questions on leadership, 20 on training, and the remaining 60 items MOS specific. Not all MOS were expected to have an MOS-specific test. Those that did not would only test on leadership and training (Department of the Army, 1994). The guidance was that MOS questions must be broadly based, testing an entire MOS rather than a specific task, duty position, or item of equipment. The SDT was linked to the Noncommissioned Officers Education System (NCOES) as "*diagnostic tests to assist school commandants and command sergeants major in assessing the leadership skills of soldiers being sent to schools*" (Fisher, 2001, p. 415). Moreover, the SDT was to be entirely self-development; units were prohibited from allocating training time for preparation for the SDT.

The SDT was administered on a trial basis in 1990-1991 to about 125,000 soldiers. Starting in October 1991, the only test administered was the SDT. [66] The plan was to administer SDT to all E5-E7 during the period of FY1992 and FY1993, during which time tests would be scored and feedback provided, but the test results would not be used for personnel actions. Starting

[65] E5 were later returned to testing.

[66] The last test with an SQT designation was administered in September 1991.

in 1994, the SDT was to be part of the EPMS system and SDT scores would be incorporated into personnel records. This did not happen. In 1991 there was a significant force reduction in TRADOC staffing and the ability of the proponents to produce tests was significantly curtailed. In December 1994, the TRADOC Commander, General William W. Hartzog, acting with the recommendation of the Command Sergeant Majors as expressed through the SMA, terminated the SDT program. The last SDT test was administered in February 1995 (Fisher, 2001). This effectively ended the U.S. Army's 36 year job competency evaluation program which had begun in January 1959.[67]

The Aftermath – Army Job Testing Reborn

The end of the SDT did not, however, end the interest in Army competency assessment. In 1999, under the personal initiative of Army Chief of Staff General Eric K. Shinseki, the Army undertook a series of studies to determine training and leader development requirements of all members of the Army—officers, noncommissioned officers, and civilians. The NCO Panel convened study groups from around the Army and incorporated input from over 33,000 NCOs. The result was *The Army Training and Leader Development Panel (ATLDP) Report (NCO)* released in April 2002 (Department of the Army, 2002). Significant in the report were the following conclusion and recommendation:

> *The Army has no method of objectively assessing NCOs' proficiency in tactical and technical MOS skills and grade-related leadership skills. A majority of NCOs believe the Army should conduct*

[67] From this point until the present day, the Army is the only service without a job evaluation testing program. The Air Force, Navy, Marine Corps and Coast Guard all have annual testing programs of enlisted job incumbents.

an annual assessment of their tactical and technical MOS skills and leadership proficiency against established standards, updated as doctrine, organizations and material change. This competency assessment would improve training and readiness by providing feedback to NCOs, units, and the Army on training and leader development program effectiveness (p. 33).

[The Army should] develop and sustain a competency assessment program for evaluating soldiers' technical and tactical proficiency in the MOS and leadership skills for their rank. Army leaders must commit to the program and link the assessment and feedback to soldier, unit, and Army readiness. Begin with skill levels 1 (specialist) and 2 (sergeant), first in divisional MOSs, then those at corps and echelons above corps. After that, continue with skill levels 3 (staff sergeant) and 4 (sergeant first class) (p. 34).

During the period 1997 to 2005, the Army entered a concentrated period of studies, analysis, plans, forecasts, and evaluations in how it could best prepare itself to meet the needs and requirements of the 21st century. Among the main issues were the requirements to acquire and develop a professional enlisted force and to define and sustain competency in that force. In 2003, the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) initiated a three-year study of the feasibility of instituting Army enlisted competency assessment. The study consisted of a needs analysis, a projected prototype program, and a cost analysis (Knapp & Campbell,

2004, 2005, 2006). The study built off the experiences of the SQT program, but was primarily based on renewed interest within the Army for building a new enlisted assessment program under 21st century conditions. The ensuing years since the end of SQT/SDT had also brought about changes in the thinking of Army NCO leadership. In 2003, SMA Jack L. Tilley (who was SMA with Army Chief of Staff General Shinseki) endorsed competency assessment and outlined a specific program with features such as web-administration, a common-core assessment, situation based assessments, and incorporation into the NCO promotion system. Although it did not occur, SGM Tilley's stated desire was to inaugurate such a program during his tenure as SMA (Campbell, Knapp, & Heffner, 2005).

During the pursuance of the ARI competency feasibly studies, one of the efforts involved detailed analysis of the SQT program and an objective assessment of its strengths and faults. The major conclusion was that the premise was sound but the execution flawed. Logistics were a major factor, and the technology available today in test development and administration might have made a significant difference in the SQT (Knapp & Campbell, 2004). By 2006, significant steps had been made in operationalizing the initiation of a competency assessment program. Ultimately, however, the increasing operational tempo pressures of Iraq and Afghanistan deployments consumed the primary attention of the Army and progress toward reinstituting competency testing abated. Whether Army interest in competency testing is renewed in the future, and what form it might take, is a chapter still to be written.

References

Army Training and Doctrine Command. (2003). *Transforming the Army. TRADOC's first thirty years 1973-2003* (TRADOC Historical Study Series). Fort Monroe, VA: Military History Office, United States Army Training and Doctrine Command.

Bell, W. G., & Cocke, K. E. (1977). *Department of the Army historical summary. Fiscal year 1973*. Washington, DC: Center of Military History. Retrieved November 9, 2008. http://www.history.army.mil/books/DAHSUM/1973

Branson, R. K. (1975). *Interservice procedures for instructional systems development* (5 vols). Fort Benning, GA: US Army Combat Arms Training Board.

Bryne, A. H. (1971). *Enlisted evaluation system*. Retrieved November 1, 2008. sill-www.army.mil/FAMAG/1971/AUG_1971/AUG_1971_PAGES_46_51.pdf

Campbell, R. C., Ford, P., & Campbell, C. H. (1978). *Development of a workshop on construction and validation of Skill Qualification Tests*. Alexandria, VA: Human Resources Research Organization.

Campbell, R. C., (1994). *The Army Skill Qualification Test (SQT) program* (IR-PD-94-05). Alexandria, VA: Human Resources Research Organization.

Campbell, R. C., Knapp, D. J., & Heffner, T. S. (2005). *Planning for the future: Progress toward an NCO competency assessment program* (Special Report No. 61). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Chapman, A. W. (1991). *The Army's training revolution 1973-1990. An overview*. Fort Monroe, VA: Office of the Command Historian, United State Army Training and Doctrine Command.

Cocke, K. E. (1995). *Department of the Army historical summary. Fiscal year 1985*. Washington, DC: Center of Military History. Retrieved November 11, 2008. http://www.history.army.mil/books/DAHSUM/1985

Cocke, K. E. (2000). *Department of the Army historical summary for fiscal year 1975*. Washington, DC: Center of Military History. Retrieved November 8, 2008. http://www.history.army.mil/books/DAHSUM/1975

Defense Finance and Accounting Service (1958). Monthly Basic Pay and Allowances Effective: 1 June 1958. (Pay table). Retrieved November 26, 2008. http://www.dfas.mil/militarypay/2006militarypaytables/militarypaypriorrates/1958.pdf

Department of Defense (1957). *Report of Defense Advisory Committee on Professional and Technical Compensation*. Washington, DC: U.S. Government Printing Office.

Department of the Army (1964). *Enlisted evaluation system* (Army Regulation No. 611-205). Washington, DC: Headquarters, Department of the Army.

Department of the Army (1977a). *Department of the Army historical summary. Fiscal year 1976*. Washington, DC: Center of Military History. Retrieved November 8, 2008. http://www.history.army.mil/books/DAHSUM/1976

Department of the Army (1977b). *Soldier's manual of common tasks* (Department of the Army Pamphlet). Fort Eustis, VA: U.S. Army Training Support Center.

Department of the Army (1978). *Enlisted personnel management system* (Army Regulation No. 600-200). Washington, DC: Department of the Army.

Department of the Army (1979). *Department of the Army historical summary. Fiscal year 1977.* Washington, DC: Center of Military History. Retrieved November 11, 2008. http://www.history.army.mil/books/DAHSUM/1977

Department of the Army (1980). *Handbook for the conduct of skill qualification tests* (Pamphlet No. 350-37). Fort Eustis, VA: US Army Training Support Center.

Department of the Army (1985). *Handbook for administration of the individual training evaluation program* (Pamphlet No. 350-37). Washington, DC: Department of the Army.

Department of the Army (1988). *Department of the Army historical summary. Fiscal year 1982.* Washington, DC: Center of Military History. Retrieved November 12, 2008. http://www.history.army.mil/books/DAHSUM/1982

Department of the Army (1994). *Leader development for America's army* (Pamphlet No. 350-58). Washington, DC: Headquarters, Department of the Army.

Department of the Army (2002). *The army training and leader development panel report (NCO)* (Final Report). Fort Leavenworth, KS: US Army Combined Arms Center and Fort Leavenworth.

Fabian, D. R. (1974). *Inside the enlisted evaluation system (EES).* Retrieved November 1, 2008. sill-www.army.mil/famag/1974/JUL_AUG_1974/JUL_AUG_1974_PAGES_54_57.pdf

Fisher, E. F. Jr. *Guardians of the republic. A history of the noncommissioned officer corps of the U.S. Army.* Mechanicsburg, PA: Stackpole Books.

Individual Training and Evaluation Directorate (1977). *Guidelines for development of skill qualification tests.* Fort Eustis, VA: U.S. Army.

Kitfield, J. (1995). *Prodigal soldiers.* Washington, DC: Potomac Books, Inc.

Knapp, D. J., & Campbell, R. C. (2004). *Army enlisted personnel competency assessment program: Phase I. Vol. 1: Needs analysis.* Arlington, VA: United States Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J., & Campbell, R. C. (2005). *Army enlisted personnel competency assessment program: Phase II report* (Tech. Rep. No. 1176). Arlington, VA: United States Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J., & Campbell, R. C. (2006). *Army enlisted personnel competency assessment program: Cost-benefit analysis and recommendations* (Contractor Report FR 05-75). Alexandria, VA: Human Resources Research Organization.

Maier, M. H., & Hirshfeld, S. F. (1978) *Criterion-referenced job proficiency testing: A large scale application* (Research Report No. 1193). Arlington, VA: Army Research Institute for the Behavioral and Social Sciences.

Oland, D. D. (1995). *Department of the Army historical summary. Fiscal year 1984.* Washington, DC: Center of Military History. Retrieved November 14, 2008. http://www.history.army.mil/books/DAHSUM/1984

United States General Accounting Office (1982). *The Army needs to modify its system for measuring individual soldier proficiency* (FPCD-82-28). Washington, DC: U.S. General Accounting Office.

United States General Accounting Office (1982, August 18). (Letter to Honorable John O. Marsh, Jr., The Secretary of the Army) Subject: "Army's Initiatives to Improve the Skill Qualification Test Program May Not Achieve Promised Changes" (GAO/FPCD-82-69). Washington, DC: U.S. General Accounting Office.

CHAPTER 10

CLASSIFICATION RESEARCH

Peter Greenston
U.S. Army Research Institute for the Behavioral and Social Sciences

Introduction

Enlistment into the military can be viewed as a two-step process: the successful applicant first passes selection screens on mental, physical, educational, and moral criteria, and then signs an enlistment contract that stipulates an accession date with initial entry training in a particular Military Occupational Specialty (MOS). The negotiation between the applicant and the guidance counselor at the Military Entrance Processing Station (MEPS), culminating in the offer and acceptance of MOS job training, constitutes the classification process in the Army. In order to be offered training in a particular MOS, the applicant must have qualifying scores on the Armed Services Vocational Aptitude Battery (ASVAB) for that training.

It is in the classification – or person-job match (PJM) – process that the Army has its first opportunity to decide how best to use the full potential of the applicant. At the present time, the Army uses a minimum standards approach that qualifies applicants if they meet the minimum ASVAB required score(s) for training in the MOS of interest. The training opportunities are presented to the applicant at the MEPS through the Recruit Quota System (REQUEST), the Army's training reservation system. The list of training opportunities is keyed to when the applicant wants to start training,

and is driven by MOS accession requirements and training schedule / seat availability. Applicants are channeled toward high priority MOS through the efforts of the guidance counselor and the use of various enlistment incentives (e.g., enlistment bonuses). The notion of putting the "right person in the right job at the right time", while viewed as desirable in principle, has in practice taken a back seat to meeting the need to fill priority jobs.

At this point we have already identified the two main avenues of classification research and development at ARI over the last 50 years: developing classification test batteries and deriving methods for efficiently using these batteries to match individuals to jobs. Classification test batteries are designed to predict either training performance or job performance, so that test battery development is logically connected to collection of performance criterion data for validation purposes. In the 1960s and 1970s reliance was upon MOS training performance (end of course grades), followed by Skill Qualification Test (SQT) program data during the mid to late 1980s that served as a defensible surrogate for job performance measures. Following the end of the SQT program in 1990, there was no readily available source for individual performance data across the Army. Thus, the current Army classification system rests on relationships identified between ASVAB and SQT scores that obtained in 1989. In addition to the need for systematic collection of new performance criterion data, the large number of MOS presents a research challenge of finding ways to generalize from representative MOS to similar ones so that the data collection can be made into a manageable effort.

Calculation of the relationship between ASVAB scores and job performance (i.e., the development of predicted performance equations) for each MOS is generally not feasible from a

statistical and cost point of view. However, if similar jobs can be formed into groups for the purpose of establishing this relationship, the effort required can be substantially reduced. Accordingly, an important research task is the formation of job families or groups of MOS which are reasonably homogeneous and for which predicted performance equations can be estimated with reasonable precision.

The ASVAB is composed of tests that are designed to relate to military performance. To the extent that certain of these relate to performance in a specified set of jobs, those tests are said to be valid with respect to that performance. To the extent that this set, or composite, of tests is more valid for one set of jobs than for other jobs, then these tests, in this context, are said to have differential validity.

Given the ongoing development of classification efficient batteries, or batteries that generate both high validity and high differential validity, the subsequent research challenge is to find better methods for classifying applicants into initial job training. As previously mentioned the Army has utilized a minimum enlistment standards approach, while ARI research has been focused on achieving maximum possible gain from the classification system—in other words, inserting optimization into the classification process. This research has gone forward as the Enlisted Personnel Allocation System (EPAS) project, in step with and dependent upon developments in the field of operations research and increases in computing power. The objective of EPAS optimization is to identify individual assignments that maximize the predicted performance of the accession cohort while meeting training management requirements.

This chapter is divided into several sections. Following this introduction, we describe in the second section the conceptual underpinnings of the classification test battery and the selection of tests for a battery, and discuss the determinants of classification efficiency and the early manual procedures for using test scores in classification. In the third section, we describe the development of operational test batteries and classification research through the early 1980s. In the fourth section, we describe the 1980s – 1990s pioneering research of Zeidner and Johnson to achieve a new level of sophistication in classification research. Zeidner and Johnson used simulation to better assess how use of particular sets of ASVAB tests for groups of jobs affected the value generated by the entire classification system. The fifth section is devoted to the development of EPAS, from its origin in 1982 as ARI Project B to the research embodied in the recent EPAS "field test" project.

Background

The pure case of selection typically concerns filling vacancies in a single job category from a list of applicants, and the selection of the best applicants for the job. The pure case of classification, in contrast, involves assignment from an applicant pool to a number of job categories. In the typical situation the number of applicants and vacancies is about the same, so the situation is characterized by interdependence of assignment, and the problem is to decide which job shall be done by which individual:

> The pure classification situation is most nearly approached in the military establishment, where a large flow of untrained youths continually pours into the organization and must be channeled into dozens of different types of specialized training and work, and where everyone who meets minimum screening

> standards must be used in some capacity (Thorndike, 1950, p. 216).

As mentioned in the introduction to this chapter, the classification enterprise involves two distinct problems. One centers on development and selection of tests which comprise the classification battery. The other concerns procedures for using the test scores to classify individuals, given a particular battery of predictors; this is the classification problem per se. The following description of classification research circa 1950 is divided into these two parts.

Selection of Tests for a Classification Battery

In the selection of tests, we first want to know what attributes each test should have if it is to be a useful member of a classification test battery, and what should be the joint characteristics of a set of tests which are to form a battery to be used for classification.

For ease of exposition[68], we consider the simplified classification problem of dividing individuals between two jobs (A&B). This permits attention to differences in aptitude for two jobs, and to a problem based on a single score representing differences in aptitude. When there are 3 or more jobs, we have a number of difference scores (A&B, A&C, B&C, ...) and are thrown back upon a larger classification problem. To reiterate, we are interested in whether an individual is more likely to be successful in job A or B; that is, we are interested in a prediction of difference in success for this individual, and we are ultimately interested in selecting for the battery the tests which best depict this difference.

[68] This section follows the discussion in Thorndike (1950), pp. 216-224.

In the most basic sense, a series of predictor or test scores for the individual are combined to yield a prediction of some measure of job success. The predicted performance equations for Job A and Job B indicate that predicted performance is a weighted sum of predictor variables. The weights are estimates of the relative importance of each predictor in predicting performance in jobs A and B.

Given our interest in identifying the best predictors for determining whether individuals are more likely to be successful in job A or B, we focus on the prediction of difference in success at job A vs. B. The ideal is to develop a battery of tests in which each test has high validity for one or two jobs but has near zero validity for the others, and in which the inter-correlations of the separate tests are low. Furthermore, in appraising a test for addition to a classification battery, Thorndike emphasized that we should be just as concerned that it has vanishing validity for some job categories as that it has high validity for others.[69]

What are the guidelines for building new tests? If we conceive of abilities (or interests or other non-cognitive predictors) as having primarily either positive or zero validity for a job, Thorndike pointed out that the fewer abilities a test taps the greater is the number of job categories for which it may have zero validity and the greater is the value it can have for purposes of classification. The classification situation seems to be one in which the simple, homogeneous test comes into its own.[70]

Is there any guidance for choosing among a pool of non-ideal existing tests: where the tests possess varying degrees of

[69] Thorndike (1950), pp. 220-222.
[70] Thorndike (1950), p. 223.

homogeneity (factorial purity), varying levels of validity, and varying patterns of content composition? Thorndike argued that any rules for selecting one particular set of tests will be very difficult to formulate, and that it will often be necessary to compromise between tests which are outstanding in differential validity and tests which are high in general validity for a wide range of jobs.[71]

A few years later, Horst (1954) extended Thorndike's work and described a technique for the development of a differential prediction battery. He presented a method for selecting from a relatively large number of tests (or "predictors") a subset that will yield the most accurate predictions of differences between all pairs of performance or outcome measures.[72] The method outlined by Horst for solving the problem proceeds by selecting one predictor at a time; the first predictor selected is the one which by itself yields the highest index of differential prediction efficiency; the second predictor selected is the one which, when combined with the first, yields the highest index of differential prediction efficiency; and so forth.

In the late 1940s and 1950s, Hubert Brogden, principal research scientist at the Army's personnel laboratory, was also investigating the classification efficiency of a test battery, and in so doing developed a theory of classification that integrates the factors affecting classification gains. He first proved the case for the use of predicted performance scores (see below) and proposed the use of mean predicted performance (MPP) as a measure of the classification efficiency of a test battery in

[71] Ibid.
[72] Horst noted that this problem is closely related to that of differential classification, which is the assignment of persons to activity groups in accordance with quotas so as to maximize the sum of all criterion indices corresponding to the job assignments.

making optimal assignments to multiple jobs (Brogden, 1955). He showed that classification gains can be approximated as a function of the number of homogenous job families, the mean predicted validity of the predictor tests, and the measured degree of interrelationships among the predictor tests; and, we note again, the higher the differential validity of a test, the lower its corresponding correlation with other tests (Brogden, 1959). The Brogden equation described the classification potential of a test battery, and also offered an analytical solution to the optimal classification problem that served as a yardstick with which to measure the iterative assignment procedures that were emerging at the time, to be followed by more sophisticated operations research techniques.

Procedures for Using Test Scores in Classification

For psychologists circa 1950 and today, the key classification concept has been that of (intra-) individual differences in aptitude: individuals whose assignments have great influence upon the aggregate outcome are those showing a wide range in potential contribution in different jobs, while the individual who makes an equal contribution in all jobs can be assigned anywhere without gain or loss to the aggregate outcome.[73]

We could assume that job performance scores are available for each individual for multiple jobs to which they could be assigned. However, this is not very likely because performance scores are difficult to collect, all the more so for multiple jobs to which an individual could be assigned. More realistically, let us assume that a test battery has been established from which predicted performance scores can be calculated for each individual across multiple measures. Note that Brogden had

[73] Thorndike (1950), p. 229.

proved that predicted performance scores for each job could be substituted for actual performance scores without changing the value of the obtained mean predicted performance.[74]

Assignments would be straightforward if it were possible to put everyone into the job with their highest predicted performance score. However, the immediate need to fill Army-critical jobs often precludes this solution. Because assignments are interdependent, optimum allocation should be done simultaneously, but that was impractical in 1950, so attention turned to establishing the best order in which individuals should be assigned to jobs. Beginning with predicted performance estimates[75] for each individual for each job category, Thorndike devised an index of expected contribution to the aggregate outcome with which to order the individuals. He proposed a measure of the spread of the individual's predicted scores (e.g. difference between highest and median score, or highest and 2nd highest, etc.). Individuals would be ordered by the spread index, followed by an attempt to assign them to their top job insofar as quotas permit. If assignments were made in this order,

[74] Of course these are not any predicted performance scores – rather, they are specified as least squares estimates of performance based on all the predictors in a battery.

[75] Thorndike (1950), pp. 224-227. Predicted performance scores are not typically comparable unless they have been measured on the same scale. To facilitate comparison across jobs, one approach is to express the performance as standard scores, for example, with mean of 50 and standard deviation of 10. Standardizing the measures to have the same mean across jobs is harmless, but imposing the same standard deviation is not because it treats the criteria as if they had been predicted with equal validity, which is typically not the case. In other words, scores on two jobs should be made comparable with respect to their probability of success. For example, if validities are .20 and .80 on jobs A and B composites, then the individual scoring 60 on each test (one standard deviation above the mean) would be given converted scores of 52 (=50+.2*10) on A and 58 (=50+.8*10) on B. These scores more closely represent the probability of success in each job.

maximum flexibility would be available for those individuals who show the greatest spread in predicted performance.

With the advent of rudimentary computing power in the late 1950s, the Army Personnel Research Office (APRO), a predecessor organization to the Army Research Institute for the Behavioral and Social Sciences (ARI), began development of an optimum computerized allocation system. According to Boldt (1964), in 1957 the U.S. Continental Army Command was concerned about the apparently lower quality of manpower being supplied to combat divisions, and asked APRO to conduct research on the allocation problem. Survey research by APRO confirmed that "the relatively low quality of men initially assigned to combat MOS was responsible for the low average aptitude level of the combat division" (Boldt, 1964, p. 2). In response, The Adjutant General (TAG) introduced a punched-card system of making assignments that utilized more information and "resulted in a substantial gain in the relative ability of enlisted men for the job to which they were assigned" (Boldt, 1964, p. 3). APRO scientists recognized that further improvements in the distribution system could be accomplished using stored program equipment (i.e. computers). They envisioned bringing together the classification framework developed by Brogden[76] and operations research techniques which had been shown as also applicable to the problem of assigning many people to many jobs. A small trial established the classification gains from an optimized system. In a large scale test on 5,000 trainees, the program successfully associated (after a 4-1/2 hour run-time) an MOS family with each trainee, and the resulting allocation surpassed that accomplished with existing procedures.

[76] Brogden had begun working on the mathematical formulation and approach to the problem of optimal allocation (Brogden, 1946).

During the mid-1960s APRO / BESRL (Behavior and Systems Research Laboratory) scientists continued development and application of the automated classification model. They conducted simulation studies, employing an approach in which individual entities were generated from a statistical profile of the applicant population. The studies investigated how classification gains varied with the formulation of Army Classification Battery (ACB) composites (see below), different mathematical methods for solving the optimization problem, and the amount and type of input information (Sorenson, 1965, 1966; Harris, 1967; Niehl & Sorenson, 1968; Johnson & Sorenson, 1971).

We return to this topic later in the chapter where we continue discussion of the development of classification optimization models.

Review of Army Classification Batteries, 1940 – Early 1980s

World War II Era, 1940-1945: The Army General Classification Test (AGCT)

By the middle of World War II, psychologists realized that new technologies and military equipment added new complexities and greater specialization to military jobs than had existed during World War I. Military psychologists saw the need to respond to these changes by creating new employment testing methods that would go beyond simple selection. They started investigating the feasibility of using the AGCT, a mechanical aptitude test, together with a clerical test for scientifically matching soldiers to military specialties. This was an important extension of the common sense approach to person-job matching spontaneously used by field commanders in World

War I, and exemplified the close association of practice and science in applied personnel psychology. [77]

The initial version (1a) of the AGCT consisted of vocabulary, arithmetic reasoning, and block counting items. There were a total of 150 items, primarily verbal items. Test norms were based on a population of Civilian Conservation Corps enrollees and soldiers (all white males), 20-29 years old (N=2,675).

The last version (3a) was introduced in 1945, and departed from previous formats. It was a battery of four tests which could be scored in total, or separately in order to provide separate measures of the abilities measured by AGCT. This was the beginning of the "classification battery" concept. The four component tests were reading and vocabulary, arithmetic reasoning, arithmetic computation, and pattern analysis. Test norms were based on a population of almost 40,000 soldiers stratified by Service command, race, age, and education.

Army Classification Battery (ACB) Era: 1945 – 1976

As early as 1941, research and operating experience indicated the need for tests to supplement the AGCT. Beginning in 1941, specific tests such as Mechanical Aptitude, Clerical Speed, Radio Code Learning, and Automotive Information were introduced at various times to supplement the AGCT in

[77] Lightfoot and Ramsberger (2000), pp. 5-6. During World War I, field commanders informally used Army Alpha and Beta selection test results to evaluate soldier strengths relative to job requirements. They were the first widely used selection instruments and the first published group intelligence tests. The Army Alpha was a written test for literate inductees who spoke fluent English. It had eight sections that covered verbal and arithmetic aptitudes, analogies, common sense, and number sequences. The Army Beta was designed for illiterate people or non-English speakers, and used pictures instead of words. The Beta included items on mazes, block counting, number similarities, and spatial perception.

classification. By the fall of 1947, ten tests were in use for classification purposes, but interpretation and appropriate use varied widely because of insufficient validity data and limited technical knowledge of classification officers. Work began on a continuing program to study various combinations of these tests which were valid for groups of Army MOS. These combinations, predictive of performance for similar MOS groups, were called "Aptitude Areas." The ACB was officially introduced in 1949 for classification, and use of the AGCT was discontinued.

Original ACB and the 1956 Reconstitution of the Aptitude Area (AA) System

At the inception of the ACB, there were ten Occupational Areas (Combat, Electronics, Electrical Maintenance, Precision Maintenance, Military Crafts, Motor Maintenance, Clerical, Graphics, General Technical, Special Assignment (including Radio Code)), and ten AA composites were introduced. The composites were each comprised of 2 to 4 tests.

The 1956 reconstitution focused on classification efficiency (Zeidner, Harper, & Karcher, 1956). When the original AA system was launched in 1949, as many AA composites were introduced as then appeared necessary to adequately differentiate the combinations of aptitudes required by the Army job structure. This approach was taken with the proviso that the effectiveness of the ACB composites would be a topic of continuing investigation as more research and data became available. The data for reconstituting the AAs were drawn from a large number of studies of enlisted men in training or in units. Scores on the ACB tests and on various composites of these tests were compared with measures of training school success and on-the-job performance.

Composites were formulated with only two tests and simply weighted. Analysis indicated that this produced as effective a composite as elaborate weighting of two or more tests. For each group of related jobs that had been studied, index values for the most promising two-test composites were listed, one measuring general ability and one specialized aptitude. The most consistently effective composite was selected for each Occupational Area: "What was desired in each case was the composite which would best predict success in the given Occupational Area and at the same time distinguish the aptitudes required in this Occupational Area from those required in other jobs – that is, would be the best differential classifier available for the Occupational Area" (Zeidner, Harper, & Karcher, 1956, p. 2).

The research indicated that a reduction in the number of AAs from 10 to 7 was in order, based on findings that only 7 composites were needed to make maximum use of the tests then in the ACB and that 7 new AAs were as effective as the original 10. The new Aptitude Areas were Combat, Electronics, General Maintenance, Motor Maintenance, Clerical, General Technical, and Radio Code. The 1956 reconstitution of the AA composites vis-à-vis its predecessor composites can be summarized as "leaner and meaner", with comparable validity and improved differential validity relative to the previous composites.

1958 New Combat Area Predictors

The AA formulation in 1949 did not include a classification measure specifically designed to meet requirements for combat. The predictors were taken from tests used for training and assignment to technical and support jobs. To remedy the situation, the Personnel Research Branch of The Adjutant General's Office (TAGO- PRB) and the Army Human

Resources Research Office (HRRO) undertook research to develop new combat predictors (see Willemin & Karcher, 1958). Researchers utilized data collected in four efforts: a 1949 study that developed and validated measures for predicting performance under arduous conditions (Arctic study); 1951 and 1953 studies conducted in Korea that identified effective and ineffective combat infantrymen using supervisory evaluations and "fighters vs. non-fighters" using peer evaluations, followed by administration of psychological tests to evaluate experimental predictor measures; and a 1955 Ft. Riley study that administered experimental tests in a garrison-maneuvers situation to confirm previous findings.[78]

Research results led to the development of two new tests. The Classification Inventory Test (CIT) asked the man to describe himself – what he has done, what he likes to do, how he sees himself as a leader – and captured important aspects of personality such as self-confidence, self-assertiveness, emotional stability, leadership, and devotion to group objectives. The General Information Test (GIT) identified masculine interests, especially outdoor activities, found to track with being an effective combat infantryman. The generation of these new tests led in turn to the modification of two combat-related composites, now designated Infantry (IN) and Artillery, Armor, and Combat Engineer (AE).

The new composites were expected to provide measures not only of what a man can do in combat but also of what he will do. The new AA composites, IN and AE, were an improvement over their predecessors because they did a better job of

[78] The Arctic study, the 1951 Korean study, and the Fort Riley study were conducted by PRB scientists; the 1953 Korean study was conducted by HRRO scientists.

identifying those with combat potential and because they overlapped less with the composites used for classification into other occupational areas (Willemin & Karcher, 1958, p. 12). The AA composites and their component ACB tests, circa 1967, are shown in Table 10.1.

Late 1960s Longitudinal Evaluation

The reconstituted AA composites described above were in place through approximately 1972. During this period, scientists at BESRL (the predecessor of ARI; see Maier & Fuchs, 1969; 1972a; 1972b) conducted a major longitudinal research effort to improve the predictive accuracy of the classification battery and corresponding AA composites in response to the needs of an increasingly technological Army. The prior system had been designed for men meeting what had been considered minimal standards of AFQT 20th percentile. At the beginning of the Vietnam War, selection standards were revised to admit men of a lower mental category (AFQT in the 10 – 15 percentile range) and the Army began drawing in more soldiers with limited general mental ability. Getting through training was problematic for these soldiers, even though they met some Aptitude Area prerequisites, because they lacked the general mental ability necessary to cope with more complicated concepts.

Table 10.1. Army Classification Battery, circa 1967

	Aptitude Area Composites							
	AE	CL	EL	GM	GT	IN	MM	RC
General Ability								
Arithmetic Reasoning					X	X		
General Information	X							
Verbal Knowledge		X			X			X
Mechanical Ability								
Electronics Information			X					
Mechanical Aptitude			X				X	
Automotive Information	X						X	
Shop Mechanics				X				
Perceptual Ability								
Pattern Analysis				X				
Radio Code Aptitude								X
Clerical Speed		X						
Self Description								
Classification Inventory						X		

Note: Aptitude Areas: AE=armor, artillery, engineers; CL=clerical; EL=electronic; GM=general maintenance; GT=general technical; IN=infantry; MM=motor maintenance; RC=radio code.

The research utilized training grades as the criterion, collecting data from the ACB and 20 experimental measures for 25,000 soldiers (all male) across 90 MOS at the start of training. MOS

were initially grouped by the then new Career Management Field (CMF), which was being considered for operational implementation in the early 1970s. CMFs were viewed as sets of closely related jobs, expected to require similar aptitudes, knowledge, and interests. Through an iterative process – starting with 39 CMFs – researchers selected those tests which best explained training performance and combined those CMFs which were "explained" by the same tests. The process of combining CMFs and dropping tests went through several cycles, and resulted in identification of nine homogenous MOS groupings and those tests with the most predictive power.

The nine MOS groups were similar to the prior system, but with several noteworthy changes (see Table 10.2). We also note that this 1972 grouping of MOS remains in place up to the present time, although the corresponding AA composites and component tests have been revised.

Each MOS group had associated with it an AA composite built from the selected tests that best predict the criteria for the MOS in that group. The composite was used in determining training eligibility for assignment to an MOS in that group. The final grouping of the MOS, the tests retained in the new ACB, and the tests selected for each composite formed the new Aptitude Area (AA) system. The 13 tests were grouped in four domains: (a) general ability, (b) mechanical ability, (c) perceptual ability, and (d) self-description (i.e., interests) inventory. The 1973 AA composites and component tests are shown in Table 10.3.

All the composites contained at least one test of general mental ability. The effect was to spread general mental ability more equitably across all occupational areas, and increase the likelihood that those accepted into training would be able to succeed. The new composites contained between three and five

tests each making them more complex than those they replaced which only contained two tests. The use of more tests per composite increased prediction accuracy of each composite by reducing the likelihood of capitalizing on chance high scores on a particular test. At the same time, though, this increased the inter-correlations between composites and likely reduced the classification efficiency of the battery.[79]

Table 10.2. Nine MOS Groups, 1972-73

Group	Relationship to Prior Grouping
Clerical (CL)	Similar
Combat (CO)	New: Infantry (prior IN) combined with Armor & Engineering (from prior AE)
Electronics Repair (EL)	Similar: attempted to separate electronics and electrical repair, but not supported empirically
Field Artillery (FA)	New: Artillery split off (from prior AE)
General Maintenance (GM)	Similar
Mechanical Maintenance (MM)	Similar: vehicle drivers split off (to new OF)
Operators & Food (OF)	New: formed around missile crewmen (from prior AE), motor vehicle drivers, and food service workers
Surveillance & Communication (SC)	New: formed with radio operators (from prior RC), communications center operators (from prior CL), combat surveillance and target acquisition (from prior AE)
Skilled Technical (ST)	New label / in place of General Technical (GT)

[79] The GT composite lost its association with an MOS group, but continued to serve as a general mental ability indicator of eligibility for additional / reclassification training. In the 1973 scheme, the GT composite score equaled the sum of Arithmetic Reasoning and Word Knowledge tests.

Table 10.3. Aptitude Area Composites and Tests, 1973

	Aptitude Area Composites								
	CL	CO	EL	FA	GM	MM	OF	SC	ST
General Ability									
Arithmetic Reasoning	X	X	X	X	X			X	X
General Information				X			X		
Mathematics Knowledge				X		X			X
Word Knowledge	X							X	
Science Knowledge					X				X
Mechanical Ability									
Trade Information		X	X			X			
Electronics Information			X	X		X			
Mechanical Comprehension			X		X			X	
Automotive Information					X	X	X		
Perceptual Ability									
Pattern Analysis		X						X	
Attention-to-Detail	X	X							
Auditory Perception								X	
Self Description									
Combat Scale		X							
Attentiveness Scale	X			X			X		
Electronics Scale			X						
Maintenance Scale						X			

The new system was estimated to be about 20 percent more effective in predicting training success than the prior system. The

average validity coefficient across composites was 0.65, compared to 0.55 for the prior system. The improved effectiveness was estimated to bring a 20 percent reduction in AIT attrition and a 20 percent reduction in the number of marginal performers.

Armed Services Vocational Aptitude Battery (ASVAB) Era: 1976 – early 1980s

Military selection and classification testing and personnel decisions in the mid-1970s were chaotic (Maier, 1993, pp. 36 – 38). In 1973, use of the common AFQT was made optional by the Office of the Secretary of Defense. The Services were allowed to derive an AFQT score from their classification batteries, with the intent that they should all be on the same score scale. In separate development efforts, these AFQT scores were linked to the original common AFQT, but not directly to each other. This left questions about a common meaning to the score scales. In addition, the separate Service batteries were logistically cumbersome, requiring applicants to take multiple batteries if they wished to be considered by more than one Service. It was during this period that "ASVAB was conceptualized as the joint-service selection and classification battery that would restore order to the accessioning process" (Maier, p. 37).

The DoD ASVAB Working Group was able to get ASVAB Forms 5/6/7 into place in time to meet a 1 January 1976 deadline. In moving toward a uniform battery, the rule adopted by the Working Group was that the tests in Forms 5/6/7 should allow each Service to calculate their existing composites. The classification batteries used by the Navy, Marine Corps, and Air Force had similar content to each other, whereas the ACB 73 had unique content in addition to that shared with the other Service batteries. Due to the unique Army content, Forms 5/6/7

were expanded over the common content to accommodate the Army composites (Maier, 1993, pp. 95-96).

ASVAB test content was adjusted in the late 1970s when Forms 8/9/10 were introduced. The new forms contained only 10 subtests compared to the inter-service ASVAB with 13 subtests. The adjustments made included (a) deleting interest inventory items due to low predictive validity; (b) deleting the General Information test because it was only used by Army and its test content overlapped verbal and technical content; (c) deleting the Attention-to-Detail test due to difficulty of accurate printing, and replacing it with Coding Speed which had been in Service classification batteries since WWII; (d) adding a Paragraph Comprehension (PC) test to improve measurement of literacy and help control coaching; and (e) adding a Numerical Operations (NO) test to ASVAB 5/6/7 to control coaching, but its speeded nature led to additional problems.[80] Content of the ASVAB remained unchanged from October 1980 to approximately January 2002.

The Army has had nine aptitude composites since 1973. As described, these composites were developed for the ACB, the predecessor to the ASVAB. Because the ACB and Forms 5/6/7 (used from 1976 until October 1980) had similar subtests, the Army aptitude composites had been intact since 1973. With Forms 8/9/10 and changes to the subtests, the composites were due for change.

A related question concerned whether the number of composites should also be changed (Maier & Grafton, 1981). The decision

[80]The Space Perception test – a part of the AFQT from 1950 onward – was deleted due to adverse impact on females, and low mean incremental validity across all specialties (and Services); its differential validity was not investigated.

was made to stay with the same number of composites, because alterations would have entailed changing the relationship between composites and individual MOS and would have required coordination among many Army agencies, at least until new forms could be validated against new training programs and new job proficiency measures (Maier & Grafton, 1981, p. 4).

The criteria used by Maier and Grafton in validation were skill training outcomes (using continuous final course grades for 1976 – 77 training events) and testing outcome measures from the recently developed Skill Qualifications Test (SQT) program, measuring MOS job knowledge .[81] In conducting the validation, criterion scores were obtained for 35 skill specialties. Subtests with the highest validity were selected for each composite; as a rule subtests were added to the composites as long as they increased predictive validity.[82] The revised composites and component subtests are shown in Table 10.4 (along with average validities).[83] Note the emphasis on predictive validity, and away from classification efficiency, in the development of these composites and battery.

[81] See Maier and Grafton (1981, pp 4-5), for a defense of the use of SQT. Performance, which is what a person actually does on the job, is a different concept than proficiency, which refers to the skills and knowledge a person possesses. The first generation SQTs were measures of proficiency rather than performance. The second generation, beginning in the 1980s, emphasizes hands-on performance more than paper and pencil tests for some specialties; the validity of the ASVAB for these criteria was not tested. Arabian and Mason (1986) reported a high correlation between written SQT scores and Project A hands-on and job knowledge tests.

[82] There were a couple of exceptions to the above procedure. EL and GM were developed through expert judgment because criterion data were not available before implementing ASVAB 8/9/10; SQT scores subsequently became available for evaluation. Electronics Repair (EL) and Clerical (CL) were constructed to be identical for all Services.

[83] Unit weights were selected for calculation of the composites.

Maier and Grafton pointed out that the changing nature of criterion measures required new strategies for validating ASVAB (Maier & Grafton, 1981, pp. 11-15). While their research was confined to traditional training programs, modern instructional technology – with pass/fail scoring and self-paced instruction – was forcing a reevaluation of the utility of ASVAB as a predictor of training success. Under these circumstances, there was growing interest among Army personnel managers in the effectiveness of the ASVAB as predictors of job proficiency and job performance.

Table 10.4. Aptitude Area Composites and Tests, ASVAB Forms 8/9/10

	Aptitude Area Composites								
	CL	CO	EL	FA	GM	MM	OF	SC	ST
Estimated Validity	.53	.56	.59	.63	.73	.52	.61	.55	.75
General Ability Tests									
Arithmetic Reasoning		X	X	X		X			
Mathematics Knowledge			X	X	X				X
Verbal (WK + PC)*	X						X	X	X
General Science			X		X				X
Mechanical Ability Tests									
Electronics Information			X		X	X			
Mechanical Comprehension		X		X		X	X		X
Auto & Shop Information		X			X	X	X	X	
Other Tests									
Numerical Operations	X						X	X	
Coding Speed	X	X		X				X	

* Verbal is the sum of Word Knowledge and Paragraph Comprehension.

The research demonstrated that ASVAB appeared to be an adequate predictor of job proficiency as well as traditional training outcomes. The validity coefficients found in predicting SQT scores were comparable to those found in predicting final course grades in traditional training programs, supporting the

feasibility of employing SQT as job proficiency criteria in classification research and validation studies. In subsequent research under Project A (discussed elsewhere) ARI did explore the adequacy of the ASVAB to predict measures of job performance (Campbell & Knapp, 2001).

Development of a "Modern" Classification Research Framework

Establishing Criterion Substitutability

An integral part of the Army's Project A was the development of measures of technical proficiency in individual MOS, labeled core technical proficiency (CTP) for nine MOS and training-based knowledge components for another six MOS. CTP consisted of hands-on performance and job knowledge components and was considered by many to be the only justifiable benchmarks of job proficiency. However, the costs of CTP tests in construction, administration, scoring and maintenance made them prohibitively expensive for use in classification research covering more than 150 MOS. Therefore the research task was to determine if more readily available (and much more voluminous) measures of job knowledge (i.e. the MOS testing data coming from the SQT program starting in the mid-1980s) could be used as a source of criterion data (Zeidner, Johnson, & Vladimirsky, 1997).

The overall objective of the ensuing investigation, conducted by Zeidner and Johnson, was to determine the adequacy of operational SQT measures of job knowledge to serve as surrogates of CTP measures in developing classification procedures using ASVAB as predictor or assignment variables. The SQT criterion would be considered adequate if it could be shown that the same developmental decisions were reached or

that equivalent findings (outcomes) were obtained for both SQT and CTP.

Zeidner and Johnson found that either criterion, SQT or CTP, selected nearly the same tests for each job family composite, provided comparable estimated weights, revealed similar patterns of test validities across ASVAB tests, and produced similar MPP scores in simulated optimal assignments. Their overall conclusion was that either criterion could serve as a surrogate for the other in conducting classification research on the ASVAB because key decisions made using either job knowledge or a job performance measure were very similar and outcomes were judged to be equivalent.[84]

The finding of criterion substitutability between the CTP benchmark and the SQT program data provided a foundation for a large body of Army aptitude composite, job family, and classification optimization research. It was this investigation that justified the utilization of SQT data as criterion measures in subsequent research throughout the 1980s, 1990s, and early 2000s.

Differential Assignment Theory and Laboratory Simulation Studies of Classification Efficiency

Joseph Zeidner and Cecil Johnson, along with several colleagues, played a central role in ARI classification research efforts in the 1980s, 1990s, and early 2000s as scientists at ARI, the Institute for Defense Analyses, and the Department of Administrative Sciences at The George Washington University. They called the conceptual basis for their research Differential Assignment Theory (DAT), emphasizing that DAT is not a content theory (because its

[84] Zeidner, Johnson, Vladimirsky (1997), p. viii.

concepts pertain to research methodology). The core of that methodology is the adoption of predicted performance as a criterion variable in the multiple job assignment context, and the acceptance of mean predicted performance (MPP) – rather than the validity of predictor composites – as the preferred measure of both selection and classification efficiency (Johnson & Zeidner, 1995, p. 21).

This approach originated with Brogden and Horst in the examination of the factors affecting the gains from classification and the development of classification efficient batteries. The DAT approach is founded on two propositions. First, DAT postulates that people and jobs can be differentiated on the basis of aptitudes and abilities, while recognizing the dominant role played by general mental ability. As a result, the focus is upon identifying those aptitudes and abilities "within" general mental ability that are required in the performance of certain jobs. DAT argues that reasonable care in selecting experimental test batteries can assure the detection of such aptitudes and abilities. Second, within the single applicant pool – multiple job context that is the Army assignment system, maximizing classification efficiency is brought about by increasing differential validity (as opposed to predictive validity).[85]

The DAT approach provided the methodology for an extensive research program. One of the main efforts by Zeidner, Johnson, and colleagues was the development of an "operational" two-tiered classification system (TTCS) with which they explored the interplay between the formulation of ASVAB composites,

[85] DAT approach is summarized in Diaz, Ingerick, and Lightfoot (2004, p. 3). DAT is consistent with the findings of the dominant role of general mental ability in predictive validity, but avoids justification of tailored test composites on the basis of incremental predictive validity alone.

the number of job families, and classification efficiency using simulation methods (Zeidner, Johnson, Vladimirsky, & Weldon, 2000a).

The research methodology was comprised of several parts. First, statistical analysis was used to estimate alternative sets of ASVAB composites. At one end were the ten operational job families, where the corresponding AA composites were the (then) existing unit-weighted composites of four ASVAB tests. This set was compared with 10, 17, and 150 job families[86], where the corresponding composites were estimated from analyses of ASVAB tests against SQT criterion data.

Simulation results by job family configuration are shown in Table 10.5. The gains from making optimal assignments are shown for configurations of 10, 17, and 150 job families using empirical weights based on statistical analyses (least-squares estimates, or LSEs) and unit-weighted AA composites.[87] As can be seen, the gains in going from unit-weighted to empirically-weighted LSE composites were approximately five-fold (at ten operational job families), while there was a 50% gain in going from 10 to 150 families. The gains depicted were called potential classification gains, inasmuch as the optimal assignment simulations were constrained only by MOS quotas and stopped short of the full bevy of training management constraints. Also note that these results refer to the 9-test ASVAB; with deletion of Coding Speed and Numerical Operations tests from the battery in January 2002, Zeidner and

[86] These were identified by clustering techniques designed to yield the most homogenous sets of job families at various levels of aggregation.
[87] The 17 job family structure was obtained by shredding the existing AA families within the boundaries of the operational classification families to maximize the Horst index of classification efficiency. The ten job families refer to the nine operational families plus the family formed using the GT (general technical) composite.

Johnson estimated that potential classification efficiency would fall by approximately six percent.[88]

Table 10.5. Comparison of Various Job Family Composites Yielding Maximally Obtainable MPP: Classification Effects Only

	MPP (9-test ASVAB)
150 job families (LSEs)	.195
17 job families (LSEs)	.145
10 operational job families (LSEs)	.123
10 operational job families (AAs)	.023

Source: Zeidner, Johnson, Vladimirsky, and Weldon (2000a), Table 12, p. 58.

The TTCS concept envisioned an automated optimization algorithm working largely unseen with the 150 first-tier job families to identify job assignments that best match applicant aptitudes and to squeeze as much classification efficiency from the system, while Army recruiting, counseling, and administrative functions would be recorded / documented at a higher (second-tier) level – the existing nine / ten operational job families – so as to avoid additional administrative problems. Zeidner and Johnson (2000a, pp. 58-9) pointed out that the estimated classification gains rival those obtained from the existing selection system, and if confirmed would represent a strong case for changes to the Army's operational classification system. We return to this topic in summarizing subsequent evaluation of the TTCS proposal and in

[88] The speeded tests were dropped from the battery in part because of the difficulty of maintaining computer-administered speeded tests and in part because of the small contribution that NO and CS made to predictive validity.

discussing development of the Enlisted Personnel Allocation System (EPAS).

A by-product of the two-tiered classification system research was the estimation of empirically-weighted composites corresponding to the nine operational job families. As discussed, these composites are empirically defensible – they refer to actual job performance (SQT) criterion data – and superior to the unit-weighted composites in terms of predictive validity. The DoD decision to reduce the ASVAB from nine to seven tests took effect in January 2002, and ARI was able to recommend and the Army adopted the empirically-weighted composites (as originally estimated by Zeidner and Johnson, and subsequently verified by HumRRO) for its nine operational job families (Diaz, Ingerick, Lightfoot, 2004a; Greenston, 2002; Zeidner, Johnson, Vladimirsky, & Weldon, 2000a; Zeidner, Johnson, Vladimirsky, & Weldon, 2000b).[89]

Zeidner and Johnson's research extended into an examination of the gender and racial fairness of the ASVAB composites, in particular the implications for moving from a nine to a seven-test battery. Zeidner and colleagues, using an approach earlier described by Cleary (1968), examined whether performance of any group was underpredicted. Underprediction would be manifested by a group performing more successfully on the job than their ASVAB score would suggest. While some evidence of underprediction was found for females and blacks, the investigators concluded that the level of underprediction was too small to have practical significance (Zeidner, Johnson, Vladimirsky, & Weldon, 2004).

[89] The empirically-derived composites are depicted in Appendix A. Relative weights – the most important subtest receives a value of 1.000 within each composite – can be compared with the predecessor unit-weighted composites shown in Table 4.

As discussed, validity coefficients are the research psychologist's preferred measure of the relationship between criterion (e.g., performance) measures and ASVAB and other predictor measures. So-called predictive validities contribute to the larger index of classification efficiency – MPP – along with differential validity and the effects of the number and homogeneity of job families. The expert panel convened by ARI in 2001 asked Zeidner and Johnson to undertake and separately report complete validity analyses, which they did for first and second-tier job family structures (Zeidner, Johnson, Vladimirsky, & Weldon, 2002). We expect these analyses to prove useful in ongoing research to develop procedures for the systematic collection of new criterion data (see below). A crucial part of such procedures is the identification of representative MOS so as to obviate the need to collect criterion data for all MOS. A data mine of predictive validities is needed to support this approach.

Evaluation of the Two-Tier Classification System Research

Following development of the TTCS, ARI invited the Human Resources Research Organization (HumRRO) to review the work by Zeidner and Johnson (Diaz, Ingerick, & Lightfoot, 2004b). The engine of the TTCS model is the empirically-derived predicted performance equation, and the central questions concerned stability and differential validity of the corresponding test composites and their effects upon classification efficiency.[90] Stability refers to the degree to which the estimated ASVAB

[90] In accordance with DAT, "empirically-estimated weights are expected to differ across jobs (i.e., exhibit differential validity) in meaningful ways that capture systematic job-to-job differences in content and performance requirements.. ... As a result, MOS-specific composites should differentially predict Soldier performance for that MOS (or job family)" (Diaz, Ingerick, & Lightfoot, 2004b, p.6).

weights can be replicated in separate investigations. Differential validity, as described earlier, refers to the degree to which composite 1 is a strong predictor of job A and not job B and the degree to which composite 1 and composite 2 (and their predicted scores) are correlated.

HumRRO first examined the composites for common sources of error that might compromise stability. They found that composite weights at the 9 and 17 job family structure can reliably be differentiated from noise overall. At the 150 job family level, however, the high frequency of composites with weak weights on several ASVAB tests at the same time indicated that true differential validity was not achievable at this level.

They next evaluated the practical effects of composite stability and job family structure on classification efficiency. To do this they conducted a simulation to model the effects of error on MPP in the estimation of the composites and its implications for optimizing classification using the proposed job families. HumRRO found sizeable overlap in the MPP estimates across the 9 and 17 job family structures, and across the 17 and 150 job family structures. They concluded that there was little empirical support for going beyond nine job families with the SQT data available for analysis, though it did appear that limited shredding of the 9 into 12 could be beneficial. Results from the HumRRO MPP analyses are shown in Table 10.6. Although HumRRO found the 150 job family structure unstable, Johnson and Zeidner research indicated that there might be an intermediate structure between 25 and 65 job families that would offer sufficient stability and classification

efficiency to make it a candidate for service as the first tier of a two-tiered classification system (Johnson & Zeidner, 2006).[91] [92]

Table 10.6. Estimated MPP and 95% Confidence Intervals

	Lower	Mean	Upper
9 job family configuration	.092	.120	.147
17 job family configuration	.113	.140	.168
150 job family configuration	.159	.187	.214

What can we take from the development of the Zeidner and Johnson two-tier system and its evaluation by HumRRO? In the first place, the notion of a two-tier system has a certain appeal because it obviates the need for administrative changes to the existing nine operational families; more importantly, the research arguably indicates that a somewhat larger number of first-tier job families might offer potential classification gains and the requisite stability in the predicted performance equations. Second, future classification research efforts will need to address the 20 year old vintage of the SQT performance data. Such efforts would benefit from systematic collection of job performance and/or training performance data to serve as

[91] The potential for improvement is reinforced with the use of unstandardized weights as recommended by Zeidner and Johnson and shown by HumRRO to yield higher MPP. The preference by HumRRO for standardized weights is due to their distributional properties, but this issue is more directly addressed with distributional constraints in the optimization algorithm.

[92] In the Johnson and Zeidner (2006) reply, they argue that the 9/17/150 comparison is apples and oranges, because the 9 and 17 are judgment clusters and the 150 is empirically defensible, and that HumRRO did not examine a valid continuum of job families for the effects of an increasing number of job families.

criterion measures in the estimation and validation of (a new generation of) predicted performance equations.

Automated Classification Optimization Models

We return now to the question of the mechanics of improving classification and assignment procedures in the Army. In the introduction to this chapter, we noted that assignments could be improved were it possible to put everyone into the job with their highest predicted performance score. We also noted that the interdependence of assignments called for optimum allocation to be done simultaneously, while operations research techniques and computing power in the early 1950s were not sufficiently advanced to make this practical.

In 1982 Project A (see Chapter 8) was inaugurated to validate the ASVAB as a measure of job performance and its usefulness as a selection and classification tool, and to develop new cognitive and non-cognitive predictor measures. At the same time ARI began development of automated assignment tools that would utilize the improved predictor measures generated by Project A to maximum effect. This research was known as Project B, and it focused on the development of the Enlisted Personnel Allocation System (EPAS). During the 1980s, ARI scientists worked independently and with the EPAS contractor team – General Research Corporation – and so it is appropriate to introduce the EPAS research by first considering the classification research undertaken by Nord, Schmitz, and others at ARI in the mid 1980s. After describing this research, we pick up the EPAS story at the beginning and follow it up to the present time.

Research Framework for Analyzing Alternative Selection and Classification Policies

Nord and Schmitz (1989) set out to estimate the performance and utility effects of alternative selection and classification policies to guide the policy decision-making process. Within the structure of nine Army job families, their comprehensive analysis sought: (i) to measure the potential Soldier performance gains that can be achieved through changes in how Aptitude Area scores are used to make assignment decisions and changes in job entry standards (so-called cut scores); (ii) to estimate the costs and benefits of these performance gains in dollar terms; and (iii) to allow a variety of policies, varying in terms of practical feasibility as well as cost, to be compared to each other in relative terms. The issues under consideration included the proper role of job standards in the military selection process, the potential payoffs to improvements in performance measurement and prediction, and the impact of implementing EPAS compared to other allocation / assignment procedures (Nord and Schmitz, p.3-2).

Eleven different assignment procedures under 1984 job entry standards were analyzed.[93] All the policies were simulated using the same random sample of 4,000-plus accessions from 1984

[93] A total of thirty-three different policies were analyzed – eleven different assignment procedures under 1984 job entry standards, under entry standards raised 5 points for all jobs, and under entry standards raised 10 points for all jobs. The eleven assignment procedures were comprised of a random assignment procedure; the current assignment procedure; two top-down procedures; the EPAS procedure featuring optimization and sequential assignment; and six procedures featuring network optimization algorithms. The latter six compared optimization on AA score, optimization on single composite predicted performance, and optimal assignment using an approximation to full least-squares prediction.

Army enlistments.[94] The EPAS procedure was distinguished from the other procedures by a full complement of training management constraints. EPAS results were in the middle on the outcome metrics. Under EPAS, the average increase over random assignment was more than twice the gain over random assignment yielded by the current assignment system.

The simulated performance gains were evaluated using a benefit-cost model. In order to translate the performance gains generated by EPAS into economic terms, Nord and Schmitz used two approaches. The first involved equating these gains into a percentage of salary, using assumptions first introduced by Schmidt and Hunter. Based on this model, EPAS provided gains over the current system of $56 million annually.[95] The second methodology is based on the concept of economic opportunity cost, and asks what it would cost to obtain the gains using current selection and assignment procedures. Using current procedures in this case means increasing the number of high quality recruits and assigning them using the current system so as to achieve the simulated gains. The result is the opportunity cost value of the simulated gains.

[94] Current policy procedures were not actually simulated. The existing assignments were used to calculate baseline 1984 average performance for 36 MOS clusters.

[95] The first benefit methodology is based on the psychological utility theory of output valuation, and leads to the construction of a net present value (NPV) model for performance valuation. The heart of the calculation is the estimated dollar value of a one standard deviation increase in performance; they used a conservative estimate of 40% of salary. This approach is a refinement of work by Brogden (1951) and further developed by other researchers. Nord and Schmitz expanded upon the traditional model by explicitly taking into account not only the gain in performance, but also the length of time over which the individual performs (i.e. expected attrition), as well as training and recruiting costs that result under alternative policies and selection ratios (Nord and Schmitz, pp. 3-37 – 3-38).

The opportunity cost estimates parallel the NPV estimates, but were higher: the estimated cost of achieving the performance gains provided by EPAS under current selection standards through the recruitment of additional high quality soldiers was $81 million annually.[96] A second finding of the study was that it may be desirable to increase enlistment standards, if the increased standards can largely be met through screening greater numbers of TSC IIIB and IV applicants. A third finding was that research that improves differential performance (compared to validity research) is likely to produce substantial net benefits.

Project B: The Development of the Enlisted Personnel Allocation System, 1982 – 1990

The purpose of Project B was to investigate techniques for improved classification, and to create a prototype system which could support real-time enlisted classification as carried out by the Recruit Quota System (REQUEST), the Army's training reservation system. The work was conducted under contract by General Research Corporation (GRC) and is summarized in its final report by Konieczny, Brown, Hutton, and Stewart (1990). ARI scientists managed the effort and worked closely with the GRC team in all aspects. Project B EPAS research and prototype development was conducted over the 1982 – 1990 period. That development is described in some detail in Appendix B.

Optimization in classification results from identifying and achieving the best person-job matches – in the sense of predicted performance or attrition or related criteria – from a process which is inherently sequential in nature. The aim of the prototype was to insert optimization into the classification process, and in so doing address the limitations of the then current (and today's)

[96] Both the NPV and opportunity cost estimates are expressed in 1984 dollars.

REQUEST. REQUEST limitations refer to design issues and implementation practices in the formulation of the REQUEST training opportunity list presented to the applicant (while negotiating MOS and training start dates with the guidance counselor). The essence of the limitation is that REQUEST does not distinguish among minimally qualified applicants – it treats them all as the same. Consequently, it cannot identify the best person or persons for a job. And even if it could, it has no way of taking into account the interdependence of assignments and coming up with a list of near optimal training opportunities for applicants as they flow through the assignment process. Moreover, there is the added consideration that assignments made from these lists must obey aggregate accession and training management constraints.

An initial prototype was designed to address REQUEST shortcomings and developed to demonstrate capabilities in three modes: policy analysis, simulation, and operational modes. The policy analysis or planning mode allows the manager to examine the long-range feasibility of alternative policies, and was accomplished by solving the allocation problem and analyzing the results. The simulation mode provides a more detailed analysis and involved simulating applicant flow using optimal guidance from the solution to the allocation problem. In its operational mode the system was intended to provide real-time support to REQUEST.

The three modes were powered by the EPAS core modules. The heart of EPAS is the Quality Allocation Module (QAM). Its function is to allocate the supply of applicants (identified by the month in which they contract) to MOS training opportunities (identified by the month in which training starts). On the supply side, applicants are categorized into supply groups by gender,

education, AFQT category, aptitude profiles, and by contract month. A forecasting module (based on recruiting contract mission) generated future estimates of these supply groups. Allocations to training start month were constrained by monthly and total accession goals, applicant quality distribution goals, and MOS training goals. The solution to the optimization problem consists of those PJM combinations that yield the largest value (e.g., total AA score as an approximation to predicted performance) for the entire cohort over the given horizon, while meeting the training management constraints and policy goals embedded in the model.

The planning procedure represents the first stage in a two stage classification procedure. "It develops a 12 month strategy that addresses overall goals and missions, and that meets all training targets and constraints defined at the supply group and MOS cluster level of detail. This strategy is then input to a detailed, sequential classification procedure to process individuals. The resulting sequential process gains 'look-ahead' intelligence of future recruiting conditions when making classification recommendations" (Konieczny, Brown, Hutton, & Stewart, 1990a, p. 25). The detailed, individual level classification occurs in the Applicant Classification Module (ACM). In conducting simulations, the applicant was associated with a particular supply group depending on his / her characteristics and ASVAB profile, the recommended PJM for that supply group was retrieved and scored, and the acceptance of a particular option from an ordered list was simulated.[97] The modules supporting the planning procedure, in combination with the ACM, constituted the simulation mode capability.

[97] The acceptance or job choice procedures were either simplistic rules or variants of random selection (Konieczny, Brown, Hutton, & Stewart, 1990b, p. C-64).

Project B EPAS in its planning and simulation modes demonstrated the ability to function independently. The project stopped short of developing an operational mode, which would have required interfacing EPAS with REQUEST. The major consideration here was how to replace and/or blend the training opportunities generated by REQUEST from use of its search window with the EPAS Optimal Guidance (EOG) generated by the optimization algorithm.[98]

PC-EPAS Project: 1994 – 1996

The Project B end product was a research prototype operating in an Army mainframe environment. Upon completion in 1990, steps towards production implementation were never undertaken. By late 1993 there was renewed interest by Army management in the benefits of better classification procedures.[99] This resulted in a PC-EPAS demonstration project with objectives of proving the feasibility of an optimizing person-job match (PJM) model based on a PC platform; demonstrating anew the concept and power of an optimizing PJM system to current Army management; and developing planning and policy analysis capabilities at ARI that would support management of the recruitment, selection, and classification process. This work was undertaken by ARI and several consultants, and is described in McWhite (1993) and Rudnik and Greenston (1995).

[98] An effort to merge the two types of training opportunity lists was subsequently undertaken as part of the EPAS "field test" project, and is described below.

[99] Peter McWhite, an operations research scientist who had worked on the GRC team for Project B, approached Zita Simutis, ARI's Technical Director, with a plan for restarting EPAS research and development. ARI was able to spark the interest of the Director of Military Personnel Management and the research got underway in FY 1994.

PC-EPAS was built from scratch using the reports and model specifications from Project B.[100] The demonstration system was developed and installed on a relatively powerful PC. It operated smoothly in the planning mode, while requiring ad hoc interventions to do simulations. In addition to "re-inventing" the EPAS prototype, new research and development was accomplished along the way.[101]

PC-EPAS in planning and simulation mode was limited by several factors. Two in particular are worth noting. In the first place, for the sake of expediency, recruit supply and training class demand were approximated using existing contracts data. Unfortunately, this approach restricted the range of recruits and training seats which were available for matching, and in so doing restricted the improvements which could be realized through optimization. Second, the applicant job choice procedure in simulation mode was simplistic: processing serially from the top of the list. This is not a completely realistic representation of the classification process, in which the guidance counselor and applicant view a succession of screens which each presents some portion of the list of available jobs. Accordingly, one of the next steps taken in the development of EPAS was to utilize independent sources of applicant/contract supply and training requirements, and to address the REQUEST baseline and job choice modeling issues.

Towards Operational EPAS Prototype: 2000 - 2004

With support of Army management the EPAS team prepared a Functional Description of how EPAS would operate

[100] The original mainframe computer code was no longer retrievable.

[101] For example, explicit DEP (Delayed Entry Program) controls were introduced which allowed the user to set maximum DEP length by applicant AFQT category; and a variety of policy analysis scenarios were examined to demonstrate the power of an optimizing PJM system.

(Greenston, et al., 1998), continued development of the prototype PC-EPAS, and planned for a non-intrusive approach to the field test, one which relied on simulation evaluation.[102]

Development of EPAS on a server platform began in 2000. BTG Inc. (which was acquired by Titan Corporation) was the prime contractor responsible for system development, and HumRRO was engaged as a subcontractor to design and conduct the field test. The coding of the EPAS optimization model was guided by an update and elaboration of the EPAS Functional Description (Greenston, et al., 2001). A user's guide to EPAS was prepared to describe the system, including data preparation, batch optimization, and simulation procedures (Sticha & Smith, 2008).

The basic structure of the allocation model in the operational prototype is similar to that in the PC-EPAS version and built in Project B. The optimization model contains four types of training management constraints: (i) monthly total accessions; (ii) annual MOS (training) requirements; (iii) monthly MOS class seats; and (iv) monthly fill requirements for 25 priority MOS. The (linear programming) model is solved for the optimal allocations of

[102] The work included a closer look at the interface considerations between REQUEST and EPAS (McWhite and Greenston, 1997). This issue was recognized in the Project B final report, which hypothesized that the overlap between REQUEST training opportunities (displayed within its date-of-availability window) and EPAS opportunities would be limited. This would have the effect of making implementation of the EOG – as a new factor in REQUEST – problematic. McWhite subsequently examined the functioning of REQUEST MOS controls and made a case for disconnecting certain REQUEST DEP tables (that control which MOS are shown to applicants), which would have the effect of increasing the observed overlap with the EOG opportunities. While the Functional Description recognized the importance of modifying REQUEST (see FD Appendix G), it did not get into details such as proposed by McWhite. The field test subsequently conducted did not include (simulated) modifications to REQUEST controls and so an effective interface remained as an issue.

supply group and contract month to MOS and training start date. Reduced cost solutions are also obtained as a by-product; these represent a series of next best solutions, and together with the optimal solution form what is called EPAS Optimal Guidance (EOG).[103]

The new system was characterized by increased flexibility in key model parameters and methodology improvements. The number of supply groups was designed to be expandable. On the training demand side, the system was expanded and simplified to accommodate individual MOS rather than MOS clusters, and can be further expanded. With regard to the underlying number of job families (or unique predicted performance equations), the model was based on the existing nine Army job families and the corresponding regression-based AA composites. However, the number of job families can be increased as needed.[104] Finally, with the increased computing power of the operational prototype, it was possible to relax the

[103] The objective function, or the value to be optimized, is the total value of the training fill over the planning horizon. The fill is described by an allocation of supply group *i* in contract month *j* to MOS *m* in training start month *k*. The value of an allocation is given by the "performance value" of the connection between supply group and MOS. In the model run for the field test (described below), this value is an average AA score of each supply group on the governing composite for each MOS. There are 91 supply groups and 150 MOS in the current version. Applicant supply data are derived from USAREC gross contract missions, broken down by applicant demographic, quality and education groups, and clustered into supply groups based on ASVAB profiles. The accessions and fill requirements / seats data are taken from the monthly Target Report produced by EPMD / Accessions Management Branch.

[104] When greater differentiation becomes defensible and warranted (see discussion of the Two Tier Classification System), the insertion of new predicted performance equations for revised job families (following the research needed to reformulate them) can be accommodated.

restriction on the number of allowable connections between supply groups and MOS.[105]

The EPAS allocation model represents the classification process, and the classification process is the link between accession and training management. The model captures the interplay between applicant supply, accession and MOS training requirements, and training seat availability, and provides planning capabilities useful for Human Resources Command and Recruiting Command. The planning process is requirements driven, and the model can be used to investigate how the (expected) supply of contractees can be rearranged into an accession flow that maximizes predicted performance, or attrition reduction, or some other criterion while meeting accession and training requirements goals. In so doing it can be used to investigate a variety of related questions – What are the implications of changes in class schedules and training seats, of changes in MOS quality distribution goals, of the introduction of new MOS, of changes to size and quality of applicant supply, of new training management policies for meeting requirements, time in the Future Soldier Program (formerly known as the Delayed Entry Program), and average performance?

[105] In the original formulation connections between supply groups and MOS were a priori disallowed when supply group mean AA scores fell below MOS cut scores. In the alternative formulation, each supply group was represented with a truncated mean – the mean AA score of those members with scores above the cut score – and connections between all supply groups and MOS were a priori allowed. Elimination of the application of cut scores at the supply group level led to more MOS being available for assignment to each supply group with increased filtering at the applicant level. And using truncated means yields supply group "cost values" in the optimization that are more representative of the predicted performance of applicants in the MOS that would be coming from the supply group (Sticha, Diaz, et al., 2007, p.9).

EPAS Field Test, 2005 - 2006

As discussed, the ARI team took a non-intrusive simulation approach to the EPAS field test covering the FY 2002 period.[106] The research objective of the field test was to measure classification gains of EPAS as implemented using merge rules which, in fact, accorded considerable deference to the REQUEST list (Sticha, Diaz, Greenston, & McWhite, 2007). The "fixed" design that was used for the field test effectively tweaked the edges of the REQUEST list and did not explore a full integration of EOG and REQUEST (Lightfoot, Diaz, & Greenston, 2003).[107]

The research was undertaken to address three sets of questions: (1) Do the merge rules for combining EPAS guidance (EOG) and REQUEST training opportunity lists produce a sufficient mix of jobs from which the applicant can choose? (2) Does the job choice model (JCM) accurately portray applicant behavior? Does it provide the requisite realism for the simulation engine? (3) Does EPAS as implemented with these merge rules yield classification gains? Does it have any adverse effects on capability to meet accession goals, especially for high priority MOS?

[106] At the start of the monthly simulation cycle, the LP allocation model was run over the applicable horizon and solved to produce training recommendations (MOS and training start month – also known as EOG) specific for each supply group. In flowing applicants through the simulation, for each applicant in the current month the supply group affiliation was determined and the corresponding EOG was merged with the actual REQUEST list (of training opportunities) produced for that applicant. The applicant's choice from among the merged list was simulated using a job choice model (Diaz, Ingerick, & Sticha, 2007). The resulting MOS fill / available seats were updated. Following classification of a month's worth of applicants, the horizon of the allocation model was shifted one month and the cycle was repeated.

[107] The field test design was characterized by its use of actual REQUEST transactions data and particular merge rules. These rules placed the EOG item at the top of the merged list if it was also found in the REQUEST list; if not, it was discarded, whereas an item found in the REQUEST list and not in the EOG was retained for the merged list. As mentioned, questions about the adequacy of the likely overlap between REQUEST and the EOG were first raised in the Project B final report (Konieczny, Brown, Hutton, & Stewart, 1990a, p. 41). When McWhite examined the design considerations for the merge process, he also called for putting REQUEST in the driver's seat, but recognized the need for modifying the REQUEST Hierarchy procedures so that REQUEST and EPAS would not be working at cross purposes in generating a merged list (pp. 22-23).

On the adequacy of the merged job list, the research indicated that deference to the REQUEST list limited the impact of EPAS. The merge rules tended to eliminate the high-payoff EOG recommendations (i.e., those for which the applicant was best suited), and the REQUEST lists tended to reflect restricted visibility of a variety of jobs. With regard to the job choice process, the empirical JCM provided a defensible simulation engine for the process. However, the research did reveal that rank order of jobs on the REQUEST training list was not a major factor in job choice determination. Consequently, the re-ordering of the training opportunities list proved to be an ineffective way of utilizing the EOG that was visible. Finally, while priority MOS and other accession goals were met, no classification gains were realized for the "EPAS as implemented" over the REQUEST condition – identical mean AA scores were found in both simulation runs.

The lesson learned from the field test was that EPAS' strengths cannot be realized if current REQUEST constraints are not modified or over-ridden. Within a simulation approach, this means that the research must "get inside REQUEST" in order to directly impact the REQUEST list. The fixed design would give way to a dynamic design, in which REQUEST is modeled (rather than taken as fixed) albeit with a simplified set of

business rules since allowing too much complication could sidetrack the purpose of the effort, which is to capture the essence of REQUEST Hierarchy and Future Soldier Program control procedures and to explore how these controls could be modified / relaxed to facilitate utilization of EOG.[108]

References

Arabian, J. M., & Mason, J. K. (1986). *Relationship of SQT scores to Project A measures.* Paper presented at the annual conference of the Military Testing Association, American Psychological Association, Mystic, CT.

Boldt, R. F. (1964). *Development of an optimum computerized allocation system* (Tech. Research Rep. No. 1135). Washington, DC: U.S. Army Personnel Research Office.

Brogden, H. E. (1946). An approach to the problem of differential prediction. *Psychometrika, 11,* 139-154.

Brogden, H. E. (1951). Increased efficiency of selection resulting from replacement of single predictor with several differential predictors. *Educational and Psychological Measurement, 11,* 173 – 195.

[108] In addition to the design issue, the ARI team also identified several technical improvements that would likely improve EPAS classification efficiency. First, increasing the number of supply groups and improving the methods of clustering would improve the granularity of the optimization and the resulting EOG, and will have an indirect effect on classification efficiency. Second, a switch from standardized to unstandardized Aptitude Area composites (as discussed in connection with the Two Tier Classification System) would make use of all the variability in the underlying ASVAB scores and reduce the intra-individual AA score compression. Third, research by Zeidner and Johnson found that classification efficiency could be improved with an increase in the number of job families. Subsequent research undertaken by HumRRO called into question the stability of the resulting predicted performance equations for increases from 9 to 17, and to 150, but would seem to support modest increases from 9 to 12 or 13.

Brogden, H. E. (1955). Least squares estimates and optimal classification. *Psychometrika, 20,* 249-252.

Brogden, H. E. (1959). Efficiency of classification as a function of number of jobs, percent rejected, and the validity and intercorrelation of job performance estimates. *Educational and Psychological Measurement, 19,* 181-190.

Campbell, J. P., & Knapp, D. J. (Eds.) (2001). *Exploring the limits in personnel selection and classification.* Mahwah, NJ: Erlbaum.

Cleary, T. A. (1968). Test bias: Prediction of grades for Negro and white students in integrated colleges. *Journal of Educational Measurement, 5,* 115 – 124.

Diaz, T., Ingerick, M., & Lightfoot, M. A. (2004a). *Replication of Zeidner, Johnson, and colleagues' method for estimating Army Aptitude (AA) composites* (Study Report No. 2004-04). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Diaz, T., Ingerick, M., & Lightfoot, M. A. (2004b). *Evaluation of alternate Aptitude Area (AA) composites and job families for Army classification* (Study Note No. 2005-01). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Diaz, T., Ingerick, M., & Sticha, P. (2006). *Modeling Army applicants' job choices: The Enlisted Personnel Allocation System (EPAS) simulation job choice model (JCM)* (Study Note No. 2007-01). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Greenston, P. M., Mower, D., Walker, S. W., Lightfoot, M. A., Diaz, T. E., McWhite, P. B. et al. (2001). *Development of a personal computer-based personnel allocation system (PC-EPAS)* (Study Report No. 2002-01). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Greenston, P., Walker, S., Mower, D., McWhite, P., Donaldson, L., Lightfoot, M. A., Diaz, T., & Rudnik, R. (1998). *Enlisted Personnel Allocation System (EPAS) functional description* (Draft Study Report). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Greenston, P. M. (2002). *Proposed new Army Aptitude Area composites: A summary of research results* (Study Note No. 2002-03). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Harris, R. N. (1967). *A model sampling experiment to evaluate two methods of test selection* (Research Memorandum No. 67-2). Washington, DC: U. S. Army Behavioral Science Research Laboratory.

Horst, P. (1954). A technique for the development of a differential prediction battery. *Psychological Monographs: General and Applied, 68,* 1-31.

Johnson, C. D., & Sorenson, R. C. (1971). *Model sampling experimentation for manpower planning.* Paper presented at NATO conference on manpower planning models, Cambridge, England.

Johnson, C. D., & Zeidner, J. (1989). *The economic benefits of predicting job performance* (Paper No. P-2241). Alexandria, VA: Institute for Defense Analysis.

Johnson, C. D., & Zeidner, J. (1995). *Differential Assignment Theory sourcebook* (Research Note No. 95-43). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Johnson, C. D., & Zeidner, J. (2006). *Evaluation of alternative Aptitude Area (AA) Composites and job families for Army classification: A reply* (Study Note No. 2006-03). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Johnson, C. D., Zeidner, J., & Vladimirsky, Y. (1996). *Developing classification-efficient job families using Differential Assignment Theory techniques.* Washington, DC: The George Washington University.

Konieczny, F., Brown, G., Hutton, J., & Stewart, J. (1990a). *Enlisted Personnel Allocation System* (Final Tech. Rep. No. 1317-37-92-CR). Vienna, VA: General Research Corporation.

Konieczny, F., Brown, G., Hutton, J., & Stewart, J. (1990b). *Enlisted Personnel Allocation System: Appendixes* (Final Tech. Rep. No. 1317-38-90-CR). Vienna, VA: General Research Corporation.

Lightfoot, M. A., Diaz, T. E., & Greenston, P. M. (2003). *Off-line field test design for evaluating two approaches to person-job matching: The Army Recruit Quota System (REQUEST) and the Enlisted Personnel Allocation System (EPAS)* (Study Rep. No. 2003-08). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Lightfoot, M. A., & Ramsberger, P. F. (2000). *Matching recruits to jobs: Enlisted Personnel Allocation System* (Special Rep. No. 41). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Maier, M. H., & Fuchs, E. F. (1969). *Development of improved Aptitude Area composites for enlisted classification* (Tech. Research Rep. No. 1159). Washington, DC: U. S. Army Behavioral Science Research Laboratory.

Maier, M. H., & Fuchs, E. F. (1972). *Development and evaluation of a new ACB and Aptitude Area system* (Tech. Research Note No. 239). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Maier, M. H., & Fuchs, E. F. (1972). *An improved differential Army classification system* (Tech. Research Rep. No. 1177). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Maier, M. H., & Grafton, F. C. (1981). *Aptitude composites for ASVAB 8, 9, and 10* (Research Rep. No. 1308). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Maier, M. H. (1993). *Military aptitude testing: The past fifty years* (TR-93-007). Monterey, CA: Defense Manpower Data Center, Personnel Testing Division.

McWhite, P., & Greenston, P. M. (1998). *Design considerations for the Enlisted Personnel Allocation System (EPAS) in its interface with Army Recruit Quota System (REQUEST)* (Study Note No. 98-03). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Niehl, E. & Sorenson, R. C. (1968). *Simpo-1 entity model for determining the qualitative impact of personnel policies* (Tech. Research Note No. 193). Washington, DC: U.S. Army Behavioral Science Research Laboratory.

Nord, R. D., & Schmitz, E. J. (1989). Estimating performance and utilization effects of alternative selection and classification policies. In C. D. Johnson & J. Zeidner (Ed.), *The economic benefits of predicting job performance* (Paper No. P-2241). Alexandria, VA: Institute for Defense Analyses.

Rudnik, R. A., & Greenston, P. M. (1995). *Development of an Army prototype PC-based Enlisted Personnel Allocation System.* (Study Report No. 96-03). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Sorenson, R. C. (1965). *Optimal allocation of enlisted men – full regression equations vs. Aptitude Area scores* (Tech. Research Note No. 163). Washington, DC: U. S. Army Personnel Research Office.

Sorenson, R. C. (1966). *Research involving personnel resource allocation models* (Paper prepared for Conference on Automation and Personnel Administration, George Washington University). Washington, DC: U. S. Army Personnel Research Office.

Sticha, P. J., Diaz, T. E., Greenston, P. M., & McWhite, P. B. (2007). *Enlisted Personnel Allocation System (EPAS) enhancements to the Recruit Quota System (REQUEST)* (Tech. Rep. No. 1212). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Sticha, P. J., & Smith, G. W. (2008). *User guide to the Enlisted Personnel Allocation System (EPAS) software* (FR08-19). Alexandria, VA: Human Resources Research Organization.

Thorndike, R. L. (1950). The problem of classification of personnel. *Psychometrika, 15,* 215-235.

Willemin, L. P., & Karcher, K., Jr. (1958). *Development of combat Aptitude Areas* (Tech. Research Bulletin No. 1110). Washington, DC: Department of the Army, Personnel Research and Procedures Division.

Zeidner, J., Harper, B. P., & Karcher, E. K. (1956). *Reconstitution of the Aptitude Areas* (PRB Tech. Research Rep. No. 1095). Washington, DC: Adjutant General's Office (Army).

Zeidner, J., Johnson, C., Vladimirsky, Y., & Weldon, S. (1997). *The substitutability of criteria in the development and evaluation of ASVAB classification procedures* (Tech. Rep. No. 1071). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Zeidner, J., Johnson, C., Vladimirsky, Y., & Weldon, S. (2000a). *Specifications for an operational two-tiered classification system for the Army. Vol. I: Report* (Tech. Rep. No. 1108). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Zeidner, J., Johnson, C., Vladimirsky, Y., & Weldon, S. (2000b). *Specifications for an operational two-tiered classification system for the Army. Vol. II: Appendices* (Research Note No. 2000-12). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Zeidner, J., Johnson, C., Vladimirsky, Y., & Weldon, S. (2000c). *The effect of reducing the number of tests in the Armed Services Vocational Aptitude Battery (ASVAB)* (Study Note No. 2001-01). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Zeidner, J., Johnson, C., Vladimirsky, Y., & Weldon, S. (2002). *Determining composite validity coefficients for Army jobs and job families* (Study Note No. 2003-02). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Zeidner, J., Johnson, C., Vladimirsky, Y., & Weldon, S. (2004). *Fairness of Army ASVAB test composites for MOS and job families* (Study Note No. 2004-04). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

APPENDIX A:

ASVAB SUBTEST (RELATIVE) WEIGHTS COMPRISING THE AA COMPOSITES (JAN 2002)

	ASVAB Subtests						
	AR	MK	VE	AS	EI	GS	MC
AA COMPOSITES							
Clerical	1.000	.767	.980	.110	.110	.000	.148
Combat	.532	1.000	.529	.733	.343	.313	.595
Electronics Repair	.818	.890	1.000	.754	.598	.151	.469
Field Artillery	.715	1.000	.586	.673	.297	.249	.700
General Maintenance	.828	.794	.417	1.000	.577	.411	.503
Mechanical Maintenance	.339	.289	.237	1.000	.340	.060	.394
Operators / Food	.962	.600	.714	1.000	.377	.251	.636
Skilled Technical	.727	.697	1.000	.357	.230	.187	.446
Surveillance/ Communications	.685	1.000	.915	.437	.551	.019	.386
General Technical	1.000		1.000				
AFQT	.500	.500	1.000				

AR = Arithmetic Reasoning, MK = Mathematics Knowledge, VE = Verbal Knowledge, AS = Auto & Shop Information, EI = Electronics Information, GS = General Science, MC = Mechanical Comprehension

APPENDIX B:

EPAS RESEARCH AND PROTOTYPE DEVELOPMENT

Project B EPAS research and prototype development was conducted over the 1982 – 1990 period. The initial prototype was developed on the ARI mini-computer using only 10 MOS and a limited training management constraint set, and running a network optimization algorithm. It established the basic feasibility of the approach. This was followed by development of a full-scale prototype, which included all entry MOS and additional policy constraints. This prototype was converted and installed on an IBM mainframe platform, and was tested extensively using simulations to confirm the viability of the solution methodology. Development of an operational prototype followed in order to explore the impact of running the system in an operational environment, and the system was converted to run at the Army's Information System Command – Pentagon computer facility. The user interface was reworked and a more powerful linear programming (LP) optimization was put in place.

Optimization in classification results from identifying and achieving the best person-job matches – in the sense of predicted performance or attrition or related criteria – in a process which is inherently sequential in nature. The aim of the prototype was to insert optimization into the classification process, and in so doing address the limitations of the then current (and today's) REQUEST. REQUEST limitations refer to implementation practices and design issues in the formulation of the REQUEST training opportunity list presented to the

applicant (while negotiating MOS and training start dates with the guidance counselor). The Classification Hierarchy component within REQUEST plays a central role. This is the component which calculates an MOS priority index, for each MOS and date combination within the search window (see below), to determine the desirability of each person-job match. The Hierarchy utilizes two main factors – MOS Status and Applicant Qualification – to make this calculation. Each main factor is comprised of sub-factors, and all are combined using weights and transformation functions to arrive at the index.

Consider first the issues concerning Hierarchy implementation practices. The GRC investigators found that the current settings (of weights and transformation functions) result in virtually no differentiability among alternative MOS for a given applicant. They were set in this manner because the system manager did not (and still does not) have the means to define factors, weights, and functions "which will result in increased differentiability while ensuring that training management goals are met. The large number of possible combinations of weights and functions makes it virtually impossible for managers to effectively manipulate them to achieve desired results" (Konieczny, Brown, Hutton, & Stewart, 1990a, p. 13). The consequences of this are that common values were established for all combinations, and the possibility for MOS differentiability by applicant (type) was lost.[109] Unique values were established only in special cases and implemented in extreme fashion.[110] A second implementation issue concerned

[109] This is exacerbated by the practice of utilizing a large overall weight for the MOS status factor and a small overall weight for the Applicant Qualification factor, applied uniformly across all MOS.

[110] For example, 11X OSUT was given highest priority ensuring that minimally qualified personnel will always be offered 11X training if a vacancy exists.

achieving policy goals (e.g., quality distribution goals). Since the formulation did not (and does not) include factors which target these goals, Army managers must monitor MOS quality and externally set additional controls.

Turn now to the Hierarchy design issues and the sequential nature of the classification process. REQUEST does not contain look-ahead functionality, and processes each applicant independently, as if he / she is the only, last, and best person for a particular assignment. It does not consider the possibility that the applicant at hand is only minimally qualified to fill a particular training seat, and that a future applicant may be a better candidate to fill the seat. Another design issue concerned the search window methodology used to narrow the training opportunities considered. The window starts at the applicant's date of availability and opens for a predetermined length (i.e., usually 4 or 5 weeks). Use of this window necessarily restricts the training opportunities presented to the applicant. A final design issue was (and is) the inability of the system to examine the impact of policy alternatives. While it does support the guidance counselors, it provides no support to those managers making the policies followed by the counselors.

Based on the EPAS functional requirements analysis, the prototype was designed to address REQUEST shortcomings and developed to demonstrate capabilities in three modes: policy analysis, simulation, and operational modes. The policy analysis or planning mode allows the manager to examine the long-range feasibility of alternative policies, and was accomplished by solving the allocation problem and analyzing the results. The simulation mode provides a more detailed analysis and involved simulating applicant flow using optimal guidance from the

solution to the allocation problem. In its operational mode the system was intended to provide real-time support to REQUEST.

The three modes were powered by the EPAS core modules. The heart of EPAS is the Quality Allocation Module (QAM). "The QAM presented a unique challenge to Project B, for optimization is not feasible for allocating individuals on a one-by-one basis, while sequential classification cannot take into account the overall goals and missions of the Army by "looking ahead" at future applicant supply and MOS requirements. Research was performed to identify and develop techniques by which optimal strategies could be applied to an inherently sequential process" (p. 25).

The QAM module represents the allocation problem and its solution using LP optimization to obtain recommended "person-job" matches. The overall aim was to allocate the supply of applicants, identified by the month in which they contract, to MOS training opportunities, identified by the month in which the training starts. On the supply side, applicants were categorized into 81 supply groups and twelve contract months. The demand-for-training side was represented by 58 MOS clusters and 24 accession (i.e., training start) months. The LP solution consists of those PJM combinations that yield the largest objective function value (i.e., total AA score as an approximation to predicted performance, etc) for the entire cohort over the given horizon, while meeting the training management constraints and policy goals embedded in the model.

Supply groups and MOS clusters were needed in order to reduce the size of the LP optimization problem. The Quality Forecasting Module (QFM) represents the supply side. External supply forecasts and / or contract missions can be utilized.

Applicants were categorized into supply groups by gender, education, AFQT category, and ASVAB profiles. The objective function works to maximize the fit of those supply groups – measured by their ASVAB profile – with the AA requirements of the MOS clusters. The Training Requirements Module (TRM) represents the demand side. Allocations to MOS cluster training start months were constrained to obey monthly total accession goals and annual MOS cluster training requirements, as well as applicant quality and education distribution goals across MOS clusters. The QAM, QFM, and TRM modules constituted the planning mode capability.

The QAM represents the first stage in a two stage classification procedure. "It develops a 12 month strategy that addresses overall goals and missions, and that meets all training targets and constraints defined at the supply group and MOS cluster level of detail. This strategy is then input to a detailed, sequential classification procedure to process individuals. The resulting sequential process gains "look-ahead" intelligence of future recruiting conditions when making classification recommendations" (p. 25). The detailed, individual level classification occurs in the Applicant Classification Module (ACM), where the EPAS recommended person-job matches are subject to three scoring routines. The first gives a Hierarchy MOS Status like score; the second gives a Hierarchy Applicant Qualification like score; and the third gives an optimization ordered-list score (what subsequently has been called EPAS optimal guidance or EOG). Analysts were given the option of using these different routines. In conducting simulations, the applicant was associated with a particular supply group depending on his / her characteristics and ASVAB profile, the recommended PJM for that supply group was retrieved and scored, and the acceptance of a particular option from an

ordered list was simulated.[111] The QAM, QFM, TRM, and ACM modules constituted the simulation mode capability.

Project B EPAS in its planning and simulation modes demonstrated the ability to function independently. The project stopped short of developing an operational mode, which would have required interfacing EPAS with REQUEST. The major consideration here was how to replace and/or blend the training opportunities generated by REQUEST from use of its search window with the EOG generated by the LP optimization.[112] In addition to addressing this consideration, the implementation of an operational EPAS model would have required construction of a REQUEST Interface Module (RIM). The RIM would pass along the ordered-list guidance for each supply group, and statistical information for determining with which supply group individual applicants would be associated.

[111] The acceptance or job choice procedures were either simplistic rules or variants of random selection (p. C-64).

[112] An effort to merge the two types of training opportunity lists was subsequently undertaken as part of the EPAS "field test" project, and is described below.

PART III

SPECIALIZED SELECTION AND CLASSIFICATION TESTS

CHAPTER 11

SELECTING ARMY OFFICERS

Michael G. Rumsey[113]
U.S. Army Research Institute for the Behavioral and Social Sciences

Why is a book about Army selection and classification predominantly about enlisted soldiers? Why is only one chapter devoted to officers? The easy answer is that there has historically been more attention paid to enlisted research on this subject than on officer research, but then one must ask why this has been the case. This question is more difficult to answer.

The first explanation involves the difference between the screening processes for the officer and enlisted systems. Enlisted screening is basically a one-shot process. A candidate is evaluated based on that evaluation and the individual's occupational interests, he or she is selected into the Army and assigned a job classification.

Officer screening, on the other hand, takes place in stages. It is a long tradition in the United States that military officers receive a college degree prior to obtaining their commission. As leaders in an institution that is an important part of the country's social structure, they are expected to have a broad educational background. The expense for this education is generally at least partially, and often fully, borne by the military in terms of scholarships or other support. Thus, screening into one of the

[113] The author is deeply indebted to Nikki Wooten for conducting an extensive literature search in support of this chapter.

precommissioning training programs is the first stage in the officer selection process. Since it typically involves a major financial commitment on the part of the Army, it is in some respects the main event and has been the primary focus of officer selection research. Yet it is not the only event. Following this initial screening, there are multiple opportunities to evaluate the candidate's performance prior to commissioning. If a mistake has been made, it can still be rectified.

However, that cannot be the full explanation. While it is costly to separate an enlisted Soldier who has been accepted into the Army, it can be even more costly to jettison the investment made in an officer candidate's education, depending when in the process this decision is made.

Let us consider the various potential purposes of either enlisted or officer selection and classification. One is to determine the individual's likely future contribution to the Army. This contribution depends on the individual's ability to complete training, on his or her ability to perform well on the job, and on the amount of time the individual remains in his or her job. A second purpose is to determine the job placement that is likely to produce the most value to the Army.

Enlisted research is burdened with the need to address all of these purposes. It has culminated in a test battery, the Armed Services Vocational Aptitude Battery (ASVAB), which predicts the individual's ability to complete training, to perform well on the job, and which provides useful information on which job placement will yield the most benefit to the Army. Officer research has been much less burdened by the need to predict success in training or to determine the best person-job match. Officer candidates typically bring with them scores on a standardized educational test used to qualify for college, such as

the Scholastic Aptitude Test or American College Test. Therefore, although much of the research on officer selection has focused on prediction of training success, the availability of tests that predict academic success has probably contributed to the disparity between resources spent on enlisted and officer research. Furthermore, assignment to a particular type of officer job generally does not take place until the candidate graduates from pre-commissioning training, by which time there is considerable information available to guide the classification decision. Thus, person-job match research has also received a lower priority for officers than for enlisted soldiers.

Insofar as prediction of job performance is concerned, a curious dynamic has prevailed. The leadership aspects of an Army officer's job are considered highly important, which might be expected to stimulate research on leader selection. However, the limited capability of early measures to successfully predict leadership probably contributed to skepticism that testing for this ability could be fruitful. Thus, another potential stimulus for officer selection research has been relatively quiescent.

To say that officer research has not received the same level of sustained attention as enlisted research is not to suggest that it has lacked significance or historical interest. Substantial work has been done on cognitive testing to ensure that applicants have had the capability to complete initial training. Based on interest in more elusive characteristics associated with successful leadership, there has also been exploration of such non-cognitive factors as personality and interests. These efforts have contributed much not only to operational screening processes in the Army but also to the history of selection and classification research in general.

The multiple routes to achieving an officer commission have complicated officer screening research. There are three major avenues: the United States Military Academy (USMA), the Reserve Officers' Training Corps (ROTC) and, for those who have already received a civilian education and require only military education, Officer Candidate School (OCS). Each commissioning program has spawned a separate research history. Although there are commonalities in the three histories, there are also differences due in part to historical circumstances and the unique characteristics of each program.

Initial Screening

West Point

Early History: 1802 to 1942

As the oldest and most visible institution providing a pre-commissioning educational program for Army officers, the United States Military Academy has perhaps the most colorful history. It was established at West Point, New York, in 1802 in part to rectify a need for military engineers—thus civil engineering has long been an important part, but far from the only part, of its curriculum. Much of the early history of West Point is described in a book by Ambrose (1966).

Almost from the beginning, a unique feature of West Point selection involved the prevalent use of a nomination process. Early on, prospective students often provided letters "from someone of importance in politics" (Ambrose, 1966, p. 50). During the period of 1817 to 1833, when Colonel Sylvanus Thayer served as Superintendent, new cadets were nominated by the U.S. president, with the advice of the Secretary of War. In 1903, Congress formalized the "appointment" feature of

USMA selection, with senators and congressmen being given a formal and prominent role in the appointment process.

As noted earlier in this chapter, a common theme in pre-commissioning selection is how to balance screening for educational success versus success as an officer. If a cadet cannot complete the educational program then he or she will never have a chance to be an officer. However, if the cadet will ultimately fail as an officer, then the education will, from the standpoint of value to the military, have been in vain.

Initially, West Point selection was focused on future educational success. In 1812, Congress decreed that applicants must meet minimum standards in reading, writing, and arithmetic. These standards were not stringent. In the early years, the examination was characterized as "so easy that one cadet was asked only to define a fraction, to read two and one-half lines from a history book, and to write a dictated sentence on the blackboard (Ambrose, 1966, pp.83-84)." Nevertheless, many applicants could not pass.

In 1902, the requirement of an entrance test was dropped and replaced with the requirement of evidence of high school graduation. "Within five years the high rate of dismissal among cadets admitted through the certificate scheme convinced the Board that it had made a mistake, and in 1907 it began giving its own entrance examination." (Ambrose, 1966, p. 244).

World War II to 1973

By 1942, following the United States entry into World War II, the qualification tests had progressed significantly beyond those in use in the early 1800s. Applicants without proper certification had to pass 3-to-4 hour tests in Algebra, Geometry, English, and History. A high academic failure rate, particularly

in English, mathematics, and foreign language, was observed among first-year USMA students who had qualified without taking these tests. The Personnel Research Branch of the Adjutant General's Office, later to become the Army Research Institute for the Behavioral and Social Sciences (ARI), was asked to conduct research to help identify potential failures. A new test incorporating language aptitude and math components was developed based on the value of these components in predicting academic success.

After the war, West Point began exploring prediction of leadership performance beyond academic success. Self-report paper and pencil instruments were used to measure characteristics believed important for leadership. The West Point Biographical Inventory (WPBI) was developed from a number of existing measures addressing personal history, personality and background, and administered to West Point candidates in 1947. It was linked to later ratings of these individuals as cadets, with only mildly positive results (Brogden & Burke, 1950).

A later study attempted to control this tendency, sometimes referred to as "faking." A self-description instrument initially developed for ROTC was tried out at West Point under the name West Point Personal Inventory. A forced-choice methodology had guided the development of this measure, whereby item options were written in such a way that a respondent would find it difficult to determine which choice would represent him in the most positive light. Somewhat more encouraging results were obtained with this approach than with the WPBI, with a validity of .27 for predicting West Point leadership ratings, which rose to .29 when certain items were eliminated (Brogden, Burke, & Frankfeldt, 1952). These

numbers are generally comparable to how self-report measures relate to enlisted ratings of performance.

In 1958, the shift toward a more balanced approach, which incorporated assessment of leadership potential as well as academic potential, found its way into the operational screening process. Qualification tests were replaced with a "whole person" screen, later revised in 1973. This screen consists of three components: academic measures (weighted 60%), leadership potential (30%), and physical proficiency (10%). Numerous studies have demonstrated that the USMA whole person score is a good predictor of academic performance (Brown, 1987; Davidson, 1977).

Recent Research: 1990 to Present

Although this whole person screen remains the primary screen used by West Point , additional research on self report measures has been conducted at USMA in recent decades. The success of the Assessment of Background and Life Experiences (ABLE) measure for predicting ratings of first tour enlisted performance in Project A raised the question of whether such a measure would be effective for predicting officer performance. However, there was concern that the ABLE, which included items asking individuals directly about their attitudes, opinions, and feelings, might be too subject to faking. Thus, an alternate approach was advanced which focused on items that were "historical, external, objective, first-person, and primarily verifiable, at least in principle (Mael & White, 1994, p. 296)." These items formed what was called the BioABLE. Both the ABLE and the BioABLE were found useful in predicting later ratings of performance (Mael & White,).

Shortly thereafter, USMA and the Army Research Institute collaborated on an ambitious project to examine how a variety

of individual measures collected at West Point related to leadership performance. This project became known as the Baseline Officer Longitudinal Data Set, or BOLDS. As did the Officer Prediction project, the BOLDS effort covered an extensive set of individual attributes. However, BOLDS included more modern, and arguably more sophisticated, concepts of leadership than did the Officer Prediction project. It incorporated theories regarding development of cognitive functioning, affective tendencies, and behavioral qualities. Measures of social problem solving, leadership style, experienced-based knowledge, temperament, motivation, and traditional cognitive measures such as SAT and ACT to predict evaluations of USMA cadet performance, were also included (Milan, Bourne, Zazanis, & Bartone, 2002).

The BOLDS researchers capitalized on the fact that West Point had already collected considerable information on its cadets during the admissions process and beyond. They felt that this information could be combined into scales that could potentially replace existing personality measures such as the ABLE and a similar measure used in civilian contexts, the NEO Personality Inventory (NEO-PI). Their efforts met with considerable success. They found that the archival measures they constructed did indeed provide similar results to the ABLE and NEO-PI, with all three measures relating to evaluations of cadet performance (Evans, 1997; Milan et al., 2002).

Further analyses on the BOLDS data revealed that college entrance examination scores, performance on a social judgment measure, and a measure of hardiness were all useful predictors of USMA cadet performance. Among West Point cadets, hardiness has been associated with resiliency when confronted with stress (Bartone, Snook, & Tremble, 2002).

Reserve Officers' Training Corps

Early History: Foundation to 1950

After the founding of West Point, a number of other colleges offering military education emerged between 1819 and the Civil War. In 1862, the Morrill Act advanced the spread of military education by requiring land grant colleges to offer courses in military tactics. Then, in 1916, the National Defense Act laid the foundation for the Reserve Officers' Training Corps. In that year there were ROTC units in 46 schools (Osburn, Machlin, and Loeffler, 1952). Initially, ROTC was viewed primarily as a commissioning source for reserve officers. However, as ROTC expanded, it gradually became the primary commissioning source for all officers.

ROTC at most schools has generally consisted of a Basic Course offered in the freshman and sophomore years, and an Advanced Course offered in the junior and senior years. As with West Point, stringent selection standards were slow to develop. The primary screen during World War II was the requirement of an academic average of C or equivalent prior to entry into the Advanced Course. Interviews with military staff were common (Bayroff & Goodman, 1948), although the process was not standardized. Considerable variability might have been expected between how candidates were evaluated in these interviews, although one research effort provided a clue as to what evaluators might have been assessing. Research which compared performance on a test of physical fitness with an individual's rated performance in an interview generated this conclusion: "There was a remarkably close relationship between the level of physical fitness and the evaluation made by the short interview; a score below average in physical fitness was

not found in those men rated high in officer ability (Woods, et al., 1943)."

Following World War II, the Army observed an uncomfortably high rate of academic attrition among ROTC students. This prompted a program of research to improve the screening process and to initiate it in the beginning of the first college year. The screening tests developed, known as forms of the ROTC Qualifying Examination, or RQ, were based on an existing standardized educational test used at the time, the ACE. The screening tests consisted of three quantitative and three verbal tests. These were found to relate well to both military and overall academic grades, particularly the verbal tests.

The Officer Prediction Program

In the mid-1950's, existing screening methods for ROTC were questioned, just as they were being questioned for USMA. A number of prestigious review panels, one led by the distinguished Dr. Harry Harlow and another headed by Dr. Leonard Carmichael, Secretary of the Smithsonian Institution, were convened. These panels reached the conclusion that existing selection procedures were limited, particularly in their utility for identifying combat leaders, and so were in need of revamping. Existing selection procedures focused primarily on prediction of academic success rather than on future performance as a leader, whether in combat or otherwise.

The recognition of the need for more targeted selection procedures ushered in quite probably the most ambitious officer selection and classification research program ever conducted. This effort, known as the Officer Prediction program, was documented in a number of Army Behavior and Systems Laboratory (now known as the Army Research Institute) reports (of particular note: Helme, 1972; Helme, Willemin, & Day,

1971; Helme, Willemin & Grafton, 1974; Willemin & Sadacca, 1959). It lasted well over a decade, involved the collection of massive amounts of information on thousands of individuals, and included a three-day assessment center incorporating situational exercises of unusual realism and intensity. The enormous scope of this effort was due in large part to the difficulty of differentially predicting performance across different types of officer performance. Three types of officer assignments were postulated: technical, administrative, and combat. Tests had to be developed which not only addressed attributes common to success across all of these jobs, but which were sufficiently comprehensive to identify attributes which were specific to each category of jobs, if such attributes could be identified. Similarly, performance measures to assess the effectiveness of these tests needed to cover both job requirements common to all officer jobs and requirements specific to each job category.

Painstakingly, a test development process began which ultimately resulted in the Differential Officer Battery (DOB). Information tests covering topics ranging from military tactics and technology to physical sciences were included. Self-description measures covered background, interests, and attitudes. Physical performance measures particularly well adapted to military requirements included a grenade-type throw, an endurance crawl and a two-hand coordination measure. A particularly creative type of measure presented military situations in motion-picture form in which leadership decisions were the required response. The DOB was a battery of enormous breadth, incorporating 11 separate booklets.

By 1959, an early trial version of the DOB had been developed and administered to 6,500 active duty officers. Then the final

version was developed and administered to about 4,000 officers in 1961 and 1962.

While the comprehensiveness of the tests to measure relevant officer attributes was extraordinary, the method used to assess the predictive value of these tests was groundbreaking. Conventional procedures, whereby individual performance on the test could be linked to performance on the job, were deemed insufficient. Observation of performance on a single job would not allow a determination of which tests were most predictive of performance on technical, which on administrative, and which on combat assignments. To provide such information, it was necessary to construct a series of situations where the individual's performance could be evaluated in a diverse set of tasks.

These situations were combined into an Officer Evaluation Center (OEC) at Fort McClellan, Alabama. Although performance-based assessment had been used earlier by Germans and the British in the officer selection process, and by the United States to select intelligence agents in World War II (MacKinnon, 1974; Rogers, Lilley, Wellins, Fischl, & Burke, 1982), the concept was still relatively novel when the OEC was established in 1962.

A year was needed to staff the OEC and prepare it for use. Then, in 1963, those who had been tested on the DOB were administered the assessment at Fort McClellan. Approximately 900 officers attended the OEC between 1963 and 1964. Essentially, the OEC created a new world for officers , with new assignments in an unfamiliar setting over a period of three days. The examinee assumed the role of a newly assigned member of a Military Advisory Group in a friendly foreign country. His first day involved undertaking relatively routine

technical and administrative tasks, such as inspecting vehicles and correcting supply records. These responsibilities, conducted under time constraints, would formerly have been those of a higher ranking officer.

Then, at 0230 of the second day, the situation suddenly and dramatically changed. The examinee was awakened, having had just four hours sleep, and told that the host nation was at war. The remaining exercises were performed under "emergency" conditions and increasing fatigue. The examinee would be required to remain awake all of the night preceding the third testing day, and thus on the evening of the third day would begin combat operations fatigued, and preferably hungry as well.

During the second and third day, the exercises transitioned to a wartime environment. The second day, activities included evaluating captured weapons and a potential hasty airstrip, whereas on the third day, the examinee directed an evacuation of the office, prepare a march order, directed reconnaissance of a helicopter landing zone, and report on enemy activities.

These activities generated a massive amount of data, with 1,700 values provided by the DOB and over 2,000 items from the OEC assessments. The challenge of reducing such a large number of variables into a meaningful pattern proved daunting. Similar items were grouped together through a process known as factor analysis. However, the large number of variables meant that not one, but a series of factor analyses were required to determine what the major categories were. It was not until 1971 that the first of a series of major reports on the findings emerged.

The first question to be addressed was what were the major individual characteristics represented by the DOB? Seventeen

major categories were identified. The terms used to identify these categories often suggested the nature of the work to which they would be linked. The five most prominent categories were mechanical technology, combat leadership, administration, general knowledge, and outdoor activity (Helme, Willemin, & Day, 1971).

The second question was what categories of work could be identified from the OEC observations ? Eight were identified. The two most important, technical-managerial leadership and combat leadership, "were clearly differentiated as separate independent domains of behavior" (Helme, et al., 1974). Because the exercises were designed to measure behaviors in these categories, their emergence as important factors was not surprising, but their differentiation was far from a foregone conclusion. Those who viewed leadership as a single overarching form of behavior might have been surprised that the differentiation was observed, while those who viewed leadership behavior as falling into a number of distinct categories might have been surprised by the failure to yield a clear differentiation between technical and managerial performance.

When the DOB psychological factors were linked with the OEC performance factors, some clear relationships emerged. Combat leadership potential predicted leadership behavior in combat situations and intelligence staff work but not other leadership behaviors. Conversely, scientific potential and general knowledge measures predicted all the staff functions but not combat command. These findings indicated the usefulness of the DOB for differential identification of officers for combat or staff jobs (Helme, Willemin, & Grafton, 1974).

Twenty years after the Officer Prediction project, a follow-up investigation was conducted to examine how effectively OEC variables could discriminate between participants who chose to leave the Army after their initial obligation and those who remained for a full career. This follow-up was based upon an examination of official records and OEC data for the 900 OEC participants. Using summary measures from the OEC, 65% were correctly classified either into the group of officers who had left the Army after their initial 2-year commitment or the group remaining for a full 20-year career (Mays & Dyer, 1981).

At the pre-commissioning level, it was felt that the DOB could be useful as a diagnostic measure of officer potential. Thus, the process of converting the battery into an operational instrument for ROTC career counseling was begun. Clearly, the 11-booklet DOB was too long for this purpose. The most relevant measures for differential classification had to be identified and formed into a new, reduced battery. Thus, separate aptitude and interest scales of combat leadership and technical-managerial leadership were formed based on their relationship with complementary forms of performance on the OEC. Career potential and career intent scales, not designed to differentiate between combat and non-combat, were also included in this new instrument, known as the Cadet Evaluation Battery (CEB) (Rumsey & Mohr, 1978). Content contained in each CEB scale is shown in Table 11.1. The CEB was introduced as an ROTC diagnostic instrument in 1972. By 1978, the cognitive technical-managerial subtest had become part of the selection process, as an alternative for those who did not have a qualifying score on the American College Test (ACT) or the Scholastic Aptitude Test (SAT).

The introduction of the CEB coincided with the expansion of the previously all-male ROTC program to include women. Because the development of the CEB had been based on data collected entirely from males, this raised questions about its applicability to females. Thus, research was conducted to compare male and female responses on this measure. Females performed better than males on two of the three cognitive subtests, including the critically important technical-managerial subtest, while males received higher scores on all of the non-cognitive scales (Mohr & Rumsey, 1978).

The CEB was linked to later course grades in initial officer training. At the time, this training, conducted separately provided for each of the major officer branches, was known as the Officer Basic Course. The cognitive scales on the CEB were found to relate well to performance across the branches (Gilbert, 1978).

Table 11.1. Cadet Evaluation Battery Scales and Subscales[a]

Scales	Subscales
Cadet Evaluation Test	
Combat Leadership: Cognitive	Tactics
	Practical skills
Technical-Managerial Leadership: Cognitive	History, politics, and culture
	Math/Physical science
Career Potential: Cognitive	Technology operations
Cadet Evaluation Inventory	
Combat Leadership: Noncognitive	Nature endurance
	Combat engineer
	Combat leader
	Physical leader
	Nonaesthetic
	Organized sports and outdoor skills
Technical-Managerial Leadership: Noncognitive	Decisive leader
	Verbal/social leader
	Rural versus urban
	Scientific interest
	Scientific orientation
	Math/physical science interest
Career Potential: Noncognitive	Administrator noninterest
	Administrative noninterest
	Combat
	Manual versus white collar interest
Career Intent	Career intent

[a]Source: Mohr & Rumsey, 1978

RETO and the Pre-commissioning Assessment System'

Many of the significant developments in officer selection research were stimulated by major Army reviews. This was so in 1977, when an Army study group produced a 5-volume report entitled, "A Review of Education and Training for Officers (RETO)." This study group generated several recommendations for change, including the use of a "performance-based assessment program." The implementation of the RETO recommendations adhered to the philosophy that officer candidates should be chosen based on the types of jobs junior officers were required to perform. Thus, the first step was to conduct a job analysis. This analysis reviewed previous literature, and used interviews with junior officers and their supervisors to explore what these officers did and what attributes were needed to perform their jobs successfully. As a result of these interviews, instruments were developed "designed to gauge the potential of future officers in decision-making, supervisory skills, organizational skills, communication skills, and other dimensions of leadership in the military" (Rogers, Lilley, Wellins, Fischl, & Burke, 1982, p. vii).

Three different types of measures were developed: (1) situational exercises, (2) a structured interview, and (3) a paper and pencil test battery. The situational exercises were combined into an assessment center known as the Leadership Assessment Program (LAP). Although the LAP and the previously-discussed Officer Evaluation Center (OEC) shared many elements common to assessment centers, they differed dramatically in terms of purpose and implementation. The OEC was designed as a research instrument to help identify the personal characteristics needed to perform successfully as an

Army officer. The LAP was designed as an operational tool to assess individuals on an already-determined set of characteristics.

As a one-time event, the OEC was relatively unconstrained, involving exercises spread out over a 3-day period, whereas the LAP was limited to those exercises that could be administered in a 50-minute classroom period. The OEC, conducted on an Army post, presented situations relatively high in military realism. The LAP, designed to be administered to 288 individual ROTC units, consisted of more conventional exercises that, while presented as military problems, had obvious civilian counterparts. These exercises included an in-basket exercise in which the participant performed a variety of administrative tasks when assuming the role of a newly-assigned platoon leader, a leaderless group discussion, a scheduling exercise, a counseling simulation, and an oral presentation (Rogers, et al., 1982).

The Structured Interview, like the LAP, was designed to measure the skills identified in the job analysis. Questions probed for examples of previous behavior. A guide was developed to assist the interviewer in conducting the interview and evaluating applicants' responses (W. P. Burke, personal communication, March 21, 2007).

In 1981, the Precommissioning Assessment System was implemented for screening into the ROTC Advanced Course. The Structured Interview and an assessment of leadership potential were included as elements in this system along with a Physical Aptitude Exam or Physical Fitness Test, a Medical Examination, and a grade point average of 2.0 or higher. Leadership potential was to be derived from cadet or superior ratings or, in limited cases, by results from the LAP.

The third measure stimulated by the RETO recommendations was the Officer Selection Battery (OSB), built to replace the cognitive technical-managerial subtest of the CEB and had been used as the ROTC selection battery since 1976. Because so much time had elapsed since the development of the CEB, the developers of the OSB decided to take a look at the content that should be included in the new selection battery. The OSB content was based on attributes judged to be relevant to the job performance dimensions identified in the job analysis. These dimensions were: communication, interpersonal manner, administration, decision making, initiative, and technical knowledge. Because of the prominence of the combat dimension in earlier research, combat was added to this list.

The measurement of the interpersonal manner and initiative dimensions in a paper and pencil format presented a formidable challenge. The developers addressed this by utilizing scenarios representing problem situations and presenting multiple possible solutions to the examinee. The research underlying the development of the CEB was instructive for defining much of the content to be included in the technical knowledge and combat categories.

Results from ARI research successfully linked the OSB to ROTC faculty ratings and to grades in seven Officer Basic Courses, the first training provided to officers after their commissioning (Fischl, Edwards, Claudy, & Rumsey, 1986). In 1986, it was included as one of the components of the Pre-commissioning Assessment System. When the administration of the SAT or the ACT to all applicants was determined to be a feasible approach, the use of the OSB in ROTC selection was terminated. To date, the development of the OSB represents the

last major effort undertaken to develop an academically-oriented ROTC selection measure.

Within the past few years, research focused on non-cognitive precommissioning assessment has been conducted. It involves the development of construct-based measures of motivation and interests relying heavily on personal history, and builds on earlier versions of these measures yielding positive results in Special Forces and elsewhere. Chapter 2 discusses one version, the Rational Biodata Inventory (RBI), and Chapter 12 another, the Test of Adaptable Personality (TAP).

This approach has been proven successful as a predictor of attrition (Putka, 2009) and performance (Putka, Kilcullen, Legree, & Wasko, 2011) of four-year ROTC scholarship recipients (Putka, 2009), and is now used to help determine which applicants should receive such scholarships (Putka, et al., 2011).

Officer Candidate School (OCS)

Building a Foundation

Although there is currently a provision for college graduates to enter directly into OCS, the principal distinctive feature of this program is that it provides a means for enlisted soldiers to become officers. It also provides the Army extraordinary flexibility to increase or decrease its officers among its ranks. Since its modern-day beginning in 1941, OCS has produced a substantial proportion of active duty officers during times of open military conflict and has at other times decreased its output to "only a slight trickle."

The greatest emphasis on OCS selection research occurred in the period between 1941 and 1957. During that time, "nearly 50 separate experimental or operational instruments [were]

prepared and tried out," in research involving "approximately 15,000 officers" (Parrish & Drucker, 1957).

Initial interest in OCS selection was prompted by a number of perceived problems, including the failure of 18% of students to complete OCS and reports of combat leadership shortcomings of OCS graduates in the North African operations in World War II (Fryer, et al., 1948).

Early on, the need to predict both educational success and leadership performance was recognized, but the stronger foundation was available for the prediction of the former. The Army General Classification Test (AGCT), a measure of general cognitive aptitude used for selection of enlisted soldiers was found to be a good predictor of OCS grades. Its use for OCS selection began in June 1941, with a requirement that applicants score in the top third with respect to the overall population tested. Those meeting this requirement then qualified to take the Officer Candidate Test (OCT).

The OCT was composed of three content areas, interpretation of data, arithmetic reasoning, and reading comprehension, that were found to be particularly useful for predicting OCS grades. The OCT was adopted for operational use in 1942 and continued to show prediction of academic performance at levels "equal to those of the best college entrance examinations" up to 1955 (Parrish & Drucker, 1957, p. 13).

Early attempts to predict leadership performance, as opposed to academic performance, met with disappointing results. Prior to 1945, OCS leadership research was "exploratory, scattered, and tended to treat leadership in the abstract, as was typical of much leadership research in the early 1940's (Parrish & Drucker, 1957, p. 14)." A number of approaches were tried, including trait ratings and a variety of projective techniques based on German methods

of officer selection, such as free association responses to verbal stimuli, ink blots, and photographs. None of these was particularly successful (Fryer, et al., 1948).

Research begun at the end of the Second World War shifted from these broad psychological approaches to ones perceived as more relevant to the requirements of junior grade officers. Four major avenues were pursued: (1) biographical self-report, (2) interview, (3) supervisor ratings, and (4) recommendations from civilian acquaintances. Measures representing each of these were tried out on Signal Corps officer candidates at Fort Monmouth. Collectively, these instruments provided excellent prediction of personal leadership ratings by fellow students and tactical officers. The level of prediction at Fort Monmouth (Fryer, et al. 1948) was even higher (.71) than the average level of prediction (average of .60) obtained between the OCT and OCS grades (Parrish & Drucker, 1957). The measures were also examined on an Army-wide basis, with continued good success for all but the Recommendation Blank, which was dropped. The remaining instruments were adopted operationally in 1946 (Parrish & Drucker, 1957).

Integration of Women into OCS

After the formation of the Women's Army Corps in 1943, selection of applicants for male and female OCS operated on two separate tracks, with no real attempt to coordinate the two. By 1947, however, it was clear that a more parallel system was to become national policy. Thus, the Army initiated research to adapt the instruments developed for male OCS selection for use in the selection of female OCS candidates. These instruments included the biographical self-report, the interview, and ratings from superiors. Modified versions of each instrument were administered and compared with a formal evaluation ("efficiency report") of

those on active duty. The high correlation (.56) between these instruments and the operational performance measure (Markey, et al., 1949), resulted in their acceptance accepted for female OCS screening following the passage of a 1948 law that integrated women into the Regular Army.

Period of Adjustment: 1950's to Present

Although the Biographical Information Blank (BIB) introduced into the OCS screening process in 1946 subsequently underwent minor revisions, it was not until 1956 that it was replaced by a new instrument, the Officer Leadership Qualification Inventory (OLI) After research had indicated that the BIB's effectiveness had declined, items were added from other instruments that had been found to be effective predictors of leadership in other studies. These sources included instruments drawn from ROTC and West Point (Harmon, Brogden, Mechlin, & Heyman, 1952). The new items were combined with the most effective items from the OCS BIB to form the OLI. The OLI was, like the BIB, a self-description questionnaire, including interests, self-evaluations, and an "annoyance" scale—which assessed "the extent to which the applicant is annoyed by others and by situations (Parrish & Drucker, 1957, p. 4)." This battery was found to be successful in predicting both leadership and resignation from OCS. A special key was developed for prediction of attrition after it was noted in 1953 that over 20% of entering candidates were resigning in the early stages of OCS.

In 1956 the OLI, modernized versions of the interview (now called the Officer Leadership Board Interview), and the military report (now the Officer Leadership Qualification Report), were adopted operationally (Parrish & Drucker, 1957).

After the extensive research involving the DOB and the OEC led to the development of the Cadet Evaluation Battery, research was initiated to pave the way for the adoption of the CEB for OCS selection. Administration of the CEB in OCS in 1975 revealed that OCS students performed better on this test than the ROTC samples. Those with more enlisted service performed better on many of the measures they might be expected to, such as Combat Leadership, Career Orientation, Career Intent, Military Orientation, and Motivation and Drive. Following collection of data needed for the establishment of appropriate standards, the CEB was authorized for use in OCS selection in 1979. When implemented in, the CEB was first titled the Candidate Evaluation Battery and later the Officer Selection Battery (OSB). The existence of two tests used in precommissioning screening, one in ROTC and one in OCS, with the same name (Officer Selection Battery) but entirely different content, was no doubt confusing. The purpose of titling the OCS version the same as the ROTC version was to facilitate the ultimate transition of the ROTC version into OCS, but in fact the transition never took place. The CEB-based OSB remained in use in OCS selection for several years.

As noted in the section on ROTC, recent precommissioning selection research has focused on non-cognitive measurement, in an effort to ensure candidates have the motivation as well as the cognitive skills to succeed. The biodata approach that has been found to work so well in predicting ROTC success also has been to be a promising approach within the OCS context, having been found to predict performance of OCS students (Allen, Babin, Oliver, & Russell, 2011).

Precommissioning Assessment Post-Entry

Thus far, the focus of this chapter has been on initial screening into each of the major pre-commissioning programs. This is where most of the officer selection research has focused, since it is the point at which the Army incurs a substantial financial obligation. Note that ROTC constitutes a special case, inasmuch there are no formal commitments with respect to non-scholarship students in the Basic Course. For ROTC, the critical decision points are when the individual enters the Advanced Course or earns a scholarship. After initial entry, or, in the course of ROTC, after the candidate and the Army have made a mutual commitment, the primary obstacle to the student gaining a commission is his or her failure to complete the program rather than further formal screening. However, during pre-commissioning training, there is an opportunity to gather other assessments that could have value in future personnel decisions, including branch assignment. This section briefly reviews research that has been done on such assessment during pre-commissioning training.

West Point has long used the Aptitude for Service Rating (ASR), a combination of peer and supervisor ratings, to assess cadet leadership development. A review in 1973 noted that: "The major, consistent finding has been that ASR is a better predictor than any other single variable, such as General Order of Merit, Academic Order of Merit, Physical Education Grades, Tactics Grades, Conduct, and several individual academic courses" (Butler, 1973). ASRs conducted as early as 1944 were found to be effective predictors of officer performance years later (Gaylord & Burke, 1949). In 1951, the ASR was compared with other measures to predict performance of 347 West Point

graduates serving in Korean combat. Again, ASR was the most valid predictor, with validities in the range of .50.

Similarly, when peer and superior ratings were obtained in OCS, these were found to be better predictors of future performance than other measures. Final peer and tactical officer evaluations of 414 Fort Riley OCS students in 1947 and 1948 were found to be better predictors of future Officer Evaluation Reports than conduct and academic grades or physical proficiency test scores. Peer and tactical officer evaluations were also found to be better predictors of evaluations of 90 OCS graduates in a combat zone during the Korean conflict than academic grades (Parrish & Drucker, 1957).

Personality testing in ROTC began as part of the initial screening process, but later was found to be more useful as part of the "on-campus phase of the ROTC Evaluation System (Kessler & Mietus, 1976, p. 1)." As with West Point selection, prediction of academic success was judged to be an insufficient basis for pre-ROTC screening. There was also concern that those selected for ROTC have the "personality and motivational traits presumed to be important in leadership" (Osburn, et al., 1952, p. 8). In 1947 the first version of the ROTC Self-Description Blank was developed from several existing measures. These measures were referred to as biographical because some items referred to the individual's background, but other items addressed the individual's current interests and preferences. The instrument was quickly adopted as an ROTC screening tool because of the perception of the need for its immediate use. However, it was soon "suspended from operational use because of the lack of proper personnel to administer it" (Osburn, p. 27).

When the utility of this instrument for predicting future leadership performance was found to be less than desired, a more ambitious

effort resulted in a new version drawn from seven existing instruments previously used in a military context. In a third version, a large pool of items was reduced to a total of 1,622 including "600 descriptions of personality traits, 460 self-estimates of ability, 74 attitudes, and 488 interests" (Osburn, et al, 1952, p. 12).

In the development of this third version, known as the ROTC Self-Description Blank Form II, considerable attention was given to the problem of positive self-representation, or faking, described earlier regarding West Point assessments. The final items used in this form were developed to minimize this problem. This form was found to have a reasonably positive relationship with peer ratings of performance, with validities between .27 and .35. Items from the second and third versions of the Self-Description Blank were combined into the ROTC Personal Inventory, which was adopted for use in the ROTC selection for direct commissioning into the Regular Army (Kessler & Mietus, 1976).

By the early 1970s, it was felt that the ROTC Inventory was in danger of becoming obsolete and needed to be replaced. A new instrument known as the Army Adaptation Inventory was thus developed to measure two dimensions--1) motivation and drive and 2) military orientation. Military orientation was a somewhat new concept, representing three categories--professional style, military values, and career motivation. Items measuring these dimensions were administered and compared with ratings of cadet performance at ROTC Advanced Camp. Both dimensions were found to correlate reasonably well with these ratings; with correlation coefficients of .20 and .21 for motivation and drive, and .26 and .27 for military orientation (Kessler & Mietus, 1976).

Conclusions

Enlisted research reported in this book, particularly the Project A and Career Force research, give us a context for interpreting the results from the diverse officer investigations. It teaches us that there are a range of outcomes that represent overall job success, and that the personal characteristics supporting one type of outcome tend to be different than the characteristics supporting another type. One of these outcomes is job proficiency, measured by hands-on and written tests. If we want to predict job proficiency, particularly written tests of proficiency, then cognitive tests of verbal or quantitative aptitude do quite well. This is particularly the case when success is measured in terms of classroom performance. These results held true for officers much as they did for enlisted soldiers.

However, there is another type of performance which is more difficult to define and predict, but which is also a critical element of success. This type of performance has been variously described as motivational, will do, or typical performance. It reflects not just what the individual can do but what an individual is observed to do over the course of a job. It can be measured by ratings or such administrative measures as awards. The best predictor of future rated performance is present rated performance. However, there is often no basis for rating individuals who are just entering a training program. Thus, the preferred alternative is self-report measures of personality, interests, and background. Although the requirements of an officer job are clearly different than the requirements of an enlisted job, research indicates that such self-report measures work as well for officer candidates as they do for enlisted candidates. A reasonable overall conclusion is

that combining cognitive and non-cognitive measures provides a better assessment of an individual's overall future performance than either cognitive or non-cognitive measures alone, whether that individual is a prospective enlisted Soldier or officer.

Certain events in the history of officer selection and classification stand out as particularly important. Perhaps the single-most significant effort was the Officer Prediction research program. It was exemplary because of its ambitious coverage of multiple predictors of individual characteristics, its groundbreaking use of a highly realistic assessment environment, and its examination of separate types of officer assignments. In some ways, the Officer Prediction effort anticipated Project A. In fact, it exceeded Project A in terms of the realism of its performance measures, but failed to offer the diversity of such measures provided in Project A.

Another major effort was the development of personality measures that incorporated highly sophisticated techniques to counter faking. Again, this paralleled later enlisted efforts, the development of the Assessment of Individual Motivation (AIM) and the Tailored Adaptive Personality Assessment System (TAPAS).

What does the future of officer selection and classification research hold? Predictions of the future are always hazardous, but there is reason to think that recent developments with respect to personality, attitude and interest measures will stimulate further work in that area. Future research on the use of cognitive measures for predicting officer success in training may be limited, as commercial measures (e.g., SAT, ACT) are available which can do this reasonably well. However, given recent advances in faking-resistant personality and interest

measures that can predict both performance and attrition, officer research on non-cognitive measures appears to have a promising future.

References

Allen, M. T., Babin, N. E., Oliver, J. T., & Russell, T. L. (2011). Predicting leadership performance and potential in U. S. Army Officer Candidate School (OCS). In M. G. Rumsey (Chair), Predicting leader performance: Insights from Army officer research. Symposium conducted at the annual meeting of the American Psychological Association, Washington, DC.

Ambrose, S. E. (1966). *Duty, honor, country: A history of West Point*. Baltimore: The Johns Hopkins Press.

Bartone, P. T., Snook, S. A., & Tremble, T. R. (2002). Cognitive and personality predictors of leader performance in West Point cadets. *Military Psychology, 14,* 321-338.

Bayroff, A. G., & Goodman, C. H. (1948). *Study of operational ROTC units to identify criterion and predictor variables (military and civil colleges)* (Rep. No. 762). Washington, D. C.: Personnel Research Section, Personnel Research and Procedures Branch, The Adjutant General's Office (Army).

Bayroff, A. G., & Machlin, C. T. (1949). *Validation of male Officer Candidate School selection instruments: I. Validation against graduation* (Rep. No. 819). Washington, D. C.: Personnel Research Section, Personnel Research and Procedures Branch, The Adjutant General's Office (Army).

Brogden, H. E., & Burke, L. (1950). *Validation of the West Point Biographical Inventory, WPB-1, against first year Aptitude for Service ratings*. (Rep. No. 829). Washington, DC: Personnel Research Section, Personnel Research and Procedures Branch, Adjutant General's Office (Army).

Brogden, H. E., Burke, L. K., & Frankfeldt, E. (1952). *Validation of the West Point Personal Inventory* (Rep No. 882). Washington, D. C.: Personnel Research Section, Personnel Research and Procedures Branch, Adjutant General's Office (Army).

Brown, D. C. (1987). *Military officers: Commissioning sources and selection criteria* (Final Rep. No. 87-42). Alexandria, VA: Human Resources Research Organization.

Butler, R. P. (1973). *Aptitude for the Service Rating as a predictor of junior officer performance* (Rep. No. 73-027). West Point, NY: Office of the Director of Institutional Research, U. S. Military Academy.

Carroll, J. B., Winer, B. J., Campbell, J. T., & Taylor, E. K. (1949). *The integration of WAC personnel into the Regular Army* (Rep. No. 789). Personnel Research Section, Personnel Research and Procedures Branch, The Adjutant General's Office (Army).

Davidson, T. G. (1977). *CEER/ACEER as a predictor of academic grade point average* (Rep. No. 77-014). West Point, NY: Office of the Director of Institutional Research, United States Military Academy.

Department of the Army (1978). *A Review of Education and Training for Officers* (RETO). Washington, DC: Author.

Eitelberg, M. J., Laurence, J. H., & Brown, D. C. (1992). Becoming brass: Issues in the testing, recruiting, and selection of the American military officers. In B. R. Gifford, & L. C. Wing (Eds.), *Test policy in defense: Lessons from the military for education, training, and development* (pp. 79-220). Boston: Kluwer Academic Publishers.

Evans, K. L. (1997). *Estimating personality constructs from archival data* (Tech. Rep. No. 1063). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences (NTIS No. AD A328 816).

Feifel, H., Van Steenberg, N., Brogden, H., & Klieger, W. (1952). *Measurement of the interests of Army officers* (Rep. No. 973). Washington, DC: Personnel Research Section, Personnel Research and Procedures Branch, The Adjutant General's Office (Army).

Fischl, M. A., Edwards, D. S., Claudy, J. G., & Rumsey, M. G. (1986). *Development of Officer Selection Battery Forms 3 and 4* (Tech. Rep. No. 603). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Fryer, D., Wherry, R. J., Gaylord, R. H., Bayroff, A. G., Taylor, E. K., Brandt, H., & Weinstein, R. (1948). *Procedures for selection of enlisted men for officer Training* (No. 752). Washington, DC: Personnel Research Section, Personnel Research and Procedures Branch, The Adjutant General's Office (Army).

Gaylord, R. H., & Burke, L. K. (1949). *Follow-up validation of predictor instruments for West Point classes of 1944, 1945, and 1946 against 1948 ratings on Form 67-1* (Rep. No. 811). Washington, DC: Personnel Research Section, Personnel Research and Procedures Branch, The Adjutant General's Office (Army).

Gilbert, A. C. F. (1978). *Predictive utility of the Officer Evaluation Battery*. Paper presented at the meeting of the Military Testing Association, Oklahoma City.

Harmon, L. R., Brogden, R. Mechlin, C. T., & Heyman, M. (1952). *Summary report of OCS validation studies* (Research Note No. 52-3). Washington, DC: Personnel Research Section, Personnel research and Procedures Branch, The Adjutant General 's Office (Army).

Helme, W. H. (1972). *Leadership research findings as applied to the Officer Personnel Management System* (Research Problem Review No. 72-2). Arlington, VA: U. S. Army Behavior and Systems Research Laboratory.

Helme, W. H., Willemin, L. P., & Day, R. W. (1971). *Psychological factors measured in the Differential Officer Battery* (Tech. Research Rep. No. 1173). Arlington, VA: U. S. Army Behavior and Systems Research Laboratory.

Helme, W. H., Willemin, L. P., & Grafton, F. C. (1971). *Dimensions of leadership in a simulated combat situation* (Tech Research Rep. No. 1172). Alexandria, VA: U. S. Army Behavior and Systems Research Laboratory (NTIS No. AD 730 315).

Helme, W. H., Willemin, L. P., & Grafton, F. C. (1974). *Prediction of officer behavior in a simulated combat situation* (Tech. Research Rep. No.1182). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences (NTIS No. AD 779 445).

Johnson, C. D., Haggerty, H. R., King, S. E., & Klieger, W. A. (1954). *Prediction of combat effectiveness of Military Academy graduates* (Tech. Research Note No. 32). Washington, DC: Personnel Research and Procedures Division, Personnel Research Branch, The Adjutant General's Office (Army).

Katz, A., & Burke, L. K. (1955). *Prediction of combat effectiveness of Officer Candidate School graduates* (Tech. Research Note No. 50). Washington, DC: Personnel Research and Procedures Division, Personnel Research Branch, The Adjutant General's Office (Army).

Kessler, J. J., & Mietus (1976). *The Army Adaptation Inventory: Development and standardization* (Tech. Paper No. 276). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Kirkland, F. R., & Good, J. V. (1978). *An evaluation of the ROTC scholarship selection system* (Research Note No. 83-30). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Kotula, L. J., & Haggerty, H. R. (1966). *Research on the selection of officer candidates and cadets* (Tech. Research Rep. No.1146). Washington, DC: Army Behavioral Evaluation Research Laboratory.

Mackinnon, D. W. (1974). *How assessment centers got started in the United States*. Pittsburgh: Development Dimensions International.

Mael, F. A., & White, L. A. (1994). Motivated to lead: Dispositional and biographical antecedents of leadership performance. In H. F. O'Neil & M. Drillings (Eds.), *Motivation: Theory and research* (pp. 285-312). Hillsdale, NJ: Erlbaum.

Markey, S. C., Melcher, J., Machlin, C. T., & King, S. H. (1949). *Validation of WAC Selection instruments* (Rep. No. 783). Washington, D. C. Personnel Research Section, Personnel Research Branch, The Adjutant General's Office (Army).

Mays, P. V., & Dyer, F. (1981). *Twenty years later: A follow-up of the Officer Evaluation Center*. Fort Benning, GA: U. S. Army Research Institute Field Unit.

Mays-Terry, P. V., & Dyer, F. N. (1986). *Follow-up of the Officer Evaluation Center* (Research Note No. 89-27). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Milan, L. M. (2002). *Analog scales as temperament measures in the Baseline Officer Longitudinal DataSet (BOLDS)* (Tech. Rep. No. 1126). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Milan, L. M., Bourne, D. R. Jr., Zazanis, M. M., & Bartone, P. T. (2002). *Measures collected on the USMA class of 1998 as part of the Baseline Officer Longitudinal Data Set (BOLDS)*. (Tech. Rep. No. 1127). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Mohr, E. S., & Rumsey, M. G. (1978). *Cadet Evaluation Battery: A comparison of 1975 male and female scores with one another and with 1971 male scores* (Tech. Paper No. 330). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Osburn, H. G., Machlin, C. T., & Loeffler, J. C. (1952). *Improvement of tests and techniques used to select ROTC students* (Tech. Research Rep. No. 977). Washington, DC: Personnel Research Branch, Personnel Research and Procedures Division, The Adjutant General's Office (Army).

Parrish, J. A., & Drucker, A. J. (1957). *Personnel research for Officer Candidate School* (Tech. Research Rep. No. 1107). Washington, DC: Personnel Research and Procedures Division, Personnel Research Branch, The Adjutant General's Office (Army).

Personnel Research Branch (1949). *Development of procedures for classification of male Officer Candidate School graduates to Arm or Service* (Rep. No.794). Washington, DC: Author, The Adjutant General's Office (Army). (AD-B969 448).

Rogers, R. W., Lilley, L. W., Wellins, R. S., Fischl, M. A., & Burke, W. P. (1982). *Development of the precommissioning Leadership Assessment Program* (Tech. Rep. No. 560). Alexandria, VA: U. S. Army Research Institute of the Behavioral and Social Sciences.

Rumsey, M. G., & Mohr, E. S. (1978). *Male and female factors on the Cadet Evaluation Battery* (Tech. Paper No. 331). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Willemin, L. P., & Sadacca, R. (1959). *Prediction of effective officer performance* (Research Study No. 59-3). Washington, DC: Personnel Research and Procedures Division, The Adjutant General's Office (Army).

Wilson, W. R., & Raskin, E. (1945). *The development of the officer classification tests, Camp Blanding, Florida, March 1945* (Rep. No. 710). Washington, D. C.: Personnel Research Section, Personnel Research and Procedures Branch, The Adjutant General's Office (Army).

Woods, W. L., Brouha, C. C., Seltzer, C., Heath, C. W., Bock, A. V., Flumerfelt, J. M.. & Wells, F. L. (1943). *Selection of officer candidates*. Cambridge, MA: Harvard University Press.

CHAPTER 12

HISTORY OF SELECTION AND CLASSIFICATION OF ARMY SPECIAL OPERATIONS FORCES

Teresa L. Russell
Human Resources Research Organization

L. Morgan Banks
US Army Special Operations Command

Mike Sanders, Michelle Wisecarver and Bob Kilcullen
U.S. Army Research Institute for the Behavioral and Social Sciences

Background on Army Special Operations

What we now call *Special Operations* has always been a component of the US military, but historically it has been a very small component. While the Department of Defense has certain units formally designated as Special Operations units, in concept, the term *special operations forces* typically refers to small units that have proficiency in "specialized, often nonconventional combat skills that are applied with adaptability, improvisation, innovation, and self-reliance" (Joint Chiefs of Staff, 2003, p. vii). Certainly, Roger's Rangers' exploits during the French and Indian Wars would meet this concept of special operations forces (Lock, 1998), as would many of the hit and run cavalry units of the American Civil War. However, many would argue that modern

special operations got its start during World War II with the Office of Strategic Services (OSS) (Simpson, 1983, p.12-19), which formulated specific procedures for selecting and training special operations personnel.

OSS Selection

The OSS was created in 1941 to collect intelligence and conduct espionage, subversion, and psychological warfare in support of the war effort (Roosevelt, 1976). The US had very little recent experience in these areas, and consequently, the selection and training of the men and women for these missions did not get off to a very productive start (The OSS Assessment Staff, 1948). The staff recruiting for the OSS faced a number of challenges. One primary hurdle was a lack of job descriptions for the positions of interest. In addition, the shifting circumstances of the war made it nearly impossible to predict the exact nature of the activities in which the selectees would be engaged, leaving the very real possibility that those selected would end up doing things quite different than originally envisioned. These challenges led to performance problems on the ground. In response, a team of social scientists created the first psychological assessment center in the US (MacKinnon, 1980). This program was developed under the leadership of Henry Murray, along with some of the best and brightest psychologists, psychiatrists, sociologists, and even cultural anthropologists that he could put together (Banks, 1995; The OSS Assessment Staff, 1948). Not only was this program the beginning of the civilian job assessment center concept, but it formed the foundation of all Special Operations Forces (SOF) assessment and selection programs that are currently in existence.

The OSS assessment lasted three and one half days, and included psychological testing, behavioral observations of the candidates under varying situational tests, and a detailed life history interview. At the time, there were competing theoretical orientations regarding how to make selection decisions based on personality. For this reason, Murray insisted on using the term "sufficient formulation" of the personality, indicating that he wanted to describe an individual's personality only in enough detail to make the selection decision. This entailed trying to integrate the various strengths and weaknesses of individuals into the personality "as a whole." The actual framework has been described this way:

> First, the job assessment skills of personnel psychologists were used as a framework for the overall program, focusing on the concepts of outcome criteria and predictive validity. The psychiatrists brought with them their skill in interviewing and diagnosing psychopathology, and the clinical psychologists brought their arsenal of test batteries and the measurement of personality variables. (Banks, 1995, p.65)

Murray and MacKinnon (1946) reported some preliminary results on the first 300 cases as follows: "6 out of every 100 men passed by the assessment staff proved to be unsatisfactory in the theater for one reason or another....Only one of the 300 cases failed because of a neuropsychiatric condition" (p. 80)."

At the end of World War II, the OSS was disbanded, and in 1947 the Central Intelligence Agency was created to conduct many of the intelligence missions for which the OSS was originally responsible. In fact, a former OSS psychologist

(Morgan, 1957) helped establish the CIA's selection program (Morgan, 1994).

Special Forces Selection

Following the disbanding of the OSS, Special Forces was the first Army special operations unit to adopt specific psychological screening procedures. The US Army created Special Forces (SF) in 1952 (Simpson, 1983). The SF mission originally focused on conducting guerilla warfare (now referred to as unconventional warfare), and helping train indigenous personnel to defeat guerrillas. Initially, SF candidates were screened and assigned by the senior SF leadership, with no formal psychological assessment (Bank, 1986). However, research into screening instruments began almost immediately.

Abelson (1954), building on previous studies from the OSS, the Counterintelligence Corps, the Marine Corps Officer Candidate School, and the Army's OCS, found that the effectiveness of SF soldiers could be reliably predicted based on personal history and interest test data. Later, the Army's Personnel Research Office (APRO) was asked to develop a selection battery for Special Forces candidates, which was implemented in 1961. At one point in late 1961, attrition from the Special Forces Qualification Course (SFQC) had increased to approximately 70 percent. Because of this, and an increasing need for more SF soldiers, APRO was asked to, "undertake research to increase the applicant pool, to tighten up the screening procedures, and consequently to reduce attrition from the training course" (Berkhouse, 1963).

As a first step, Army researchers conducted validity studies to develop a special battery of tests, the SF Selection Battery (Berkhouse, 1963; Berkhouse & Cook, 1961; Berkhouse, Mendelson, & Cook, 1961). The experimental predictor battery

contained a variety of noncognitive, self-description inventories as well as a situational judgment test and selected Army Classification Battery (ACB) Aptitude Area Composites. The criterion measure was a composite score computed across performance tests constructed for nine critical areas:

- Weapons
- Communications
- First aid
- Survival
- Land navigation
- Demolition
- Organization and development of guerrilla forces
- Aerial resupply
- Guerrilla tactics

Validity evidence led to the selection of four measures for the final battery: (1) the Infantry Aptitude Area Composite from the ACB; (2) the Special Forces Suitability Inventory, a noncognitive measure of emotional stability or general psychological adjustment; (3) the Critical Decisions Test, a measure of risk-taking and practical judgment (where a few facts were presented with stringent time limits for deliberation); and (4) the Locations Test, a spatial orientation measure that required orienting oneself according to photographs of terrain. The four measures together yielded a multiple correlation of .63 with the performance criterion (N=216 SF trainees), .55 when corrected for shrinkage. The Special Forces Selection Battery became operational in 1961. Several noncognitive measures were later designed with the intent of supplementing the Special

Forces Selection Battery (Marder & Medland, 1964), but there is no documentation of how well these newer measures worked.

Another validation study examined the usefulness of the Special Forces Selection Battery and other measures for predicting officers' academic grades, training performance, and peer ratings (Marder & Medland, 1965). The Special Forces Selection Battery, the Special Forces Qualifying Examination (verbal and math items extracted from other officer selection instruments), and a language aptitude test showed promise for predicting academic grades and to a lesser extent, peer ratings. None of the experimental measures predicted training performance evaluations.

A new experimental battery was developed and assessed in the early 1970s (Olmstead, Caviness, Powers, Maxey, & Cleary, 1972). The battery contained the ACB, the Interest Opinion Questionnaire, Life History Inventory, Military Interest Blank, an inventory designed to assess attitudes toward SF activities, the Team-Task Motivation Questionnaire, the Cognitive Test Battery, a test of physical endurance, and a personal information form. Several of these had subtests or subscales. Criterion proficiency measures included job knowledge tests, hands-on tests, and self- and peer ratings. Based on stepwise regression results (N=100), researchers identified thirteen tests for the final battery. Several of the best predictors were cognitive; five were from the Cognitive Test Battery, and three were ACB subtests. "Fighter" scores from the life history and military interest instruments as well as a "despair" score, physical endurance, and the team task motivation score made the final battery. Although it is unclear whether the new selection battery was ever used on a wide scale, the researchers did accomplish the important goals of reanalyzing the requirements for effective SF

performance and developing training aimed at increasing the proficiency with which those requirements could be met.

Army Creates SF Assessment Center

Around the mid-1970's the Army terminated use of special batteries for SF selection, relying primarily on the Army Physical Fitness Test (APFT), Armed Services Vocational Aptitude Battery (ASVAB) General Technical (GT) score, and information available from administrative records such as training experiences for SF selection (Pleban, et al., 1988). In 1981, the Army's Training and Doctrine Command (TRADOC) reviewed the SFQC, including its use of selection criteria (George & Cassady, 1981). At the time, only one in three soldiers successfully completed SF training. Numerous training problems were identified, although the authors stated that the selection criteria were sufficient.

In 1988, based partly on the TRADOC report, and also heavily on his prior experience with other assessment programs, BG Richard Potter, at the time a colonel, initiated the restructuring of the SFQC. This included a new assessment and selection program that would not only address the problems identified in the TRADOC report, but would also improve the selection process. BG Potter reportedly required his senior staff officers to read the book, *The Assessment of Men* (Office of Strategic Services, 1948), as part of their officer development.) He was concerned that the selection process was occurring during the intense training program, and felt that this created an inherent conflict between training successful soldiers and selecting out less suitable ones. His fundamental concept was to design a separate assessment and selection course, one that would be totally distinct from training. In this manner, once soldiers were determined to be suitable and were selected to attend the SFQC,

the instructors would focus exclusively on training. This should produce an attitude in the trainers of helping the students succeed, rather than critically looking at screening them out.

This program, which continues today, is called Special Forces Assessment and Selection (SFAS). In addition to the conceptual rationale for the development of SFAS, there was also a practical rationale: at the time soldiers completed a permanent change of station (PCS) move to attend SFQC. With the institution of SFAS, soldiers would be able to attend the 3-week SFAS in temporary duty (TDY) status. If soldiers were not selected for training, they would then be able to simply return to their assigned duty station. This would reduce the large quantity of soldiers left on post at Ft. Bragg, NC without positions due to failure to complete the SFQC.

At SFAS, military psychologists were assigned to the higher headquarters, the U.S. Army John F. Kennedy Special Warfare Center and School (USAJFKSWCS). In addition, personnel psychologists from the U. S. Army Research Institute consulted on the development of the SFAS program (Thompson et al, 1987; Pleban et al, 1988; Pleban, Allentoff, & Thompson, 1989), and assisted for several years in the detailed analysis of the program.

Development of paper-and-pencil and other selected predictors for SFAS involved two major steps. The first step was highly exploratory (Pleban, et al., 1988). The research team, along with the USAJFKSWCS psychologist, determined that predictors should tap the three general domains of intelligence, personality, and physical fitness. They selected measures for those domains, and compared profiles of SF and non-SF personnel on those measures. Four selected measures, the Wonderlic Personnel Test (WPT--a measure of general

cognitive ability), the Jackson Personality Inventory (JPI), the Myers-Briggs Type Indicator (MBTI), and a Biographical Questionnaire (BQ) were administered to three groups of soldiers: soldiers from the 197th Infantry Brigade (N=57), a class attending SFQC (N=339), and soldiers currently on A-Teams (N=19). The BQ contained 14 items tapping educational level, component (active-reserve), time in service, rank, specialized training received, MOS, marital status, race, and career plans. Based on practical concerns and comparisons between the samples and between students who were successful and unsuccessful in Phase I of the SFQC, they eliminated the MBTI from further consideration.

The second step was a criterion-related validation study (Pleban, Allentoff, & Thompson, 1989). The WPT, JPI, and BQ were administered to SFQC Phase I candidates. At that time, Phase I was a four-week course focusing on general subjects, teaching, leadership, patrolling, land navigation, and physical conditioning. Phase I status, the criterion, was based on six variables: (1) a map reading written exam, (2) a land navigation field exercise (FTX), (3) a confidence course, (4) a patrolling written exam, (5) a patrolling FTX, and (6) rated performance as a patrol leader. The six scores were noncompensatory; failure to reach the specified cut score on any one variable resulted in termination from SFQC. The best single predictor of Phase I status was WPT (r = .29). Four of the 16 JPI scales correlated significantly with Phase I status. Consequently, the authors recommended use of, and further research on the WPT and the four JPI scales.

The BQ items pertaining to specialized prior training were examined. Pleban et al. (1989) found that prior Ranger training was related to Phase I status; 84% of the candidates who had

graduated from Ranger school successfully completed Phase I. Reconnaissance and Jungle Warfare training also appeared to be associated with Phase I success. Analyses of the other BQ items (e.g., marital status) were not reported.

Applying Job Analysis Results for SF Selection

In the early 1990s, the Army conducted a systematic job analysis of SF jobs (Russell, Crafts, Tagliareni, McCloy, & Barkley, 1996) to gather information that would aid in the development of new SF performance measures. This goal required two types of information—a description of the job performance requirements (e.g., tasks and behaviors) of SF jobs and clarification of the individual attributes needed for effective performance. The research involved five major steps: (1) development of workshop materials and logistics, (2) administration of workshops to collect critical incidents and task and attribute ratings, (3) analysis of task and attribute data, (4) development of performance categories and behavior-based rating scales, and (5) analysis of linkages between attributes and performance categories. Active duty SF NCOs and officers and a Subject Matter Expert Panel (SMEP) composed of SF officers and NCOs at USAJFKSWCS participated in all parts of the project. The primary products of the project were behavior-based rating scales for SF jobs, definitions of individual attributes important for successful performance in SF jobs, and job task ratings. A list of attributes identified in the job analysis appears in Table 12.1. Definitions of performance dimensions common to all SF jobs appear in Table 12.2. The study identified an additional 12 performance dimensions for specific SF job skills. These formed the basis for behaviorally anchored rating scales that were later used to gather criterion ratings in validation studies.

The job analysis showed that service in SF requires soldiers to be resourceful and self-sufficient. They must be flexible enough to overcome unanticipated obstacles and adjust quickly to rapidly changing conditions without higher-order guidance. Temperament (e.g., Stress Tolerance, Achievement Orientation), in addition to intelligence and physical fitness, plays an important role in ensuring successful performance in these environments.

The Army Research Institute developed the Test of Adaptable Personality (TAP) to measure temperament characteristics important to SF performance. Temperaments were identified for measurement based on a research literature review and, more importantly, on evaluations made by experienced SF soldiers regarding the importance of various temperaments for SF job performance on deployments.

Table 12.1 Attributes for Successful Performance in SF Jobs

General Attributes

1. Judgment and Reasoning
2. Planning
3. Adaptability
4. Creativity
5. Auditory Ability
6. Mechanical Ability
7. Spatial Ability
8. Perceptual Ability
9. Basic Math
10. Advanced Math

Communication Abilities

11. Reading Ability
12. Writing Ability
13. Language Ability
14. Communication Ability
15. Non-Verbal Communication

Interpersonal Skills, Motivation, and Character

16. Persuasiveness/Diplomacy
17. Maturity
18. Autonomy
19. Team Playership
20. Dependability
21. Initiative
22. Perseverance
23. Moral Courage

Interpersonal Skills, Motivation, and Character (continued)

24. Motivating Others
25. Supervising

Physical and Psychomotor Attributes

26. Swimming
27. Physical Flexibility and Balance
28. Physical Strength
29. Physical Endurance
30. Psychomotor Ability

Interests

31. Interest in Adventure and Outdoor Activities
32. Interest in Skilled Trades
33. Interest in Other Cultures
34. Interest in People
35. Enterprising Interests

Table 12.1 Attributes for Successful Performance in SF Jobs (Continued)

Conventional Army Experiences (continued)

36. Leadership
37. Achievement and Effort
38. Personal Discipline
39. Physical Fitness and Military Bearing
40. General Soldiering Proficiency
41. Infantry Core Technical Proficiency
42. Combat Engineer Technical Proficiency

Conventional Army Experiences (continued)

43. Other Combat MOS Technical Proficiency
44. Radioteletype Operator Technical Proficiency
45. Medical Care Specialist Technical Proficiency
46. Other Non-Combat MOS Technical Proficiency

The TAP is unique in that it uses biodata questions to measure the targeted temperament constructs. This approach takes advantage of the fact that past behavior is often a good predictor of future behavior. Such construct-scored biodata tests not only measure the intended constructs (Kilcullen, White, Mumford, & Mack, 1995), they also yield criterion-related validity estimates comparable to what is achieved with traditional, empirical keyed biodata tests (Schoenfeldt, 1989; Uhlman, Reiter-Palmon, & Connelly, 1990). Moreover, construct-based biodata tests tend to produce more stable validity estimates over time (Clifton, Kilcullen, Reiter-Palmon, & Mumford, 1992; White & Kilcullen, 1992). Biodata questions focusing on observable behavior are also more resistant to deliberate faking compared to traditional personality questions

asking about general attitudes (Kilcullen, White, Mumford, & Mack, 1995).

Table 12.2. SF Job Performance Dimensions Common to All SF Jobs

A. Teaching Others. Conveying knowledge and skill to others; developing programs of instruction (POI) and tailoring material to the target audience's needs and capabilities; obtaining audience interest and involvement; presenting material in an orderly fashion; using handouts, aids, or tools; finding appropriate ways around language barriers; demonstrating own proficiency.

B. Building and Maintaining Effective Relationships with Indigenous Populations. Demonstrating respect for and engaging in behavior appropriate to indigenous culture, values, and customs; providing services and assistance to develop rapport with indigenous people and build respect for SF.

C. Handling Interpersonal Situations. Dealing with others constructively, persuading rather than forcing own way; remaining composed, even when provoked; using non-verbal communication skills to interpret behaviors; resolving disputes; allowing others to "win" confrontations.

D. Using and Enhancing Language Skills. Using foreign language skills to communicate with Host Nation/Guerilla (HN/G) or other foreign personnel; practicing and developing language skills.

E. Contributing to the Team Effort and Morale. Motivating others; communicating effectively with team members; enhancing new and existing team members' skills and readiness; building team spirit through personal interactions.

F. Showing Initiative and Extra Effort. Putting forth the effort to produce high-quality work in a timely fashion; actively pursuing self-improvement goals; volunteering for demanding tasks or extra responsibility; taking initiative; presenting a positive image of SF.

Table 12.2. SF Job Performance Dimensions Common to All SF Jobs (Continued)

G. Displaying Honesty and Integrity. Adhering to laws or rules of conduct; knowing when to put aside personal beliefs to follow policy requirements/standard operating procedures (SOPs), but taking a more difficult, morally correct course of action when appropriate; owning up to own mistakes; being truthful and genuine with others.

H. Planning and Preparing for Missions. Developing mission plans that are technically sound, well-coordinated, and likely to lead to mission accomplishment; obtaining complete information needed for planning; drawing on team member's experiences; anticipating enemy movement or other obstacles; weighing alternative courses of action; determining and preparing resources needed for mission accomplishment.

I. Decision Making. Assessing the situation and determining an appropriate course of action within a reasonable time frame; digesting information and drawing conclusions; using time, personnel, equipment, and tactics effectively; acting swiftly and decisively when needed; remaining level-headed and task-oriented in stressful situations.

J. Confronting Physical and Environmental Challenges. Defeating odds and environment to survive an ordeal; maintaining team standard of performance in physically challenging situations; preparing physically for challenge; following field survival guidance; taking steps to ensure own health and endurance.

K. Navigating in the Field. Maintaining correct direction of movement in diverse/demanding conditions; orienting self/team members using navigational aids and terrain features; noticing and taking into account map or environmental details to aid in navigating.

L. Troubleshooting and Solving Problems. Thinking of alternative ways to solve a problem; using the resources at hand to fabricate needed items; improvising from own technical knowledge of mechanical and electrical principles.

Table 12.2. SF Job Performance Dimensions Common to All SF Jobs (Continued)

M. Being Safety Conscious. Being alert to safety at all times; rigorously following safety guidelines and instructions for weapons/explosives or other hazardous materials; monitoring others to ensure compliance with SOP when using weapons/dangerous equipment; being alert to potential threat; maintaining noise/light discipline.

N. Administering First Aid and Treating Casualties. Applying emergency life-saving techniques and skills when accidents or injuries occur; treating ailments/conditions caused by the environment; following SOP for treating conditions and injuries.

Some of the temperament attributes measured by the TAP are listed in Table 12.3. The TAP also includes a 'response distortion' scale that detects and adjusts for deliberate faking on the part of the test taker.

Table 12.3 TAP Temperament Attributes

Attribute	Definition
Achievement Orientation	Working hard towards task accomplishment and giving one's best effort.
Cognitive Flexibility	Willingness to try innovative approaches for getting work done, and tolerating uncertainty and ambiguity.
Peer Leadership	Willingness to assume positions of authority and responsibility.
Fitness Motivation	Willingness to maintain a demanding exercise regimen.
Interpersonal Skills - Team Player	Willingness to work cooperatively and get along well with others.
Interpersonal Skills – Diplomat	Being extroverted and outgoing. Able to make friends easily and establish rapport with strangers.
Self Efficacy	Maintaining one's confidence and composure under stress.
Personal Discipline	Willingness to respect legitimate authority figures and to follow rules/regulations

In a concurrent validation study, Kilcullen, Mael, Goodwin, and Zazanis (1999) examined the criterion-related validity of cognitive tests, the TAP temperament scales, physical fitness indices, and various demographic characteristics for predicting the field performance of 210 SF enlisted soldiers serving in operational detachments. In addition, the relationship between performance in the SF selection courses and subsequent field performance was examined. The criterion consisted of evaluations provided by superiors of the SF soldier's on-the-job performance while deployed. The field performance of SF enlisted soldiers was best predicted by several TAP scales, including Achievement Orientation (r = .37), Fitness Motivation (r = .20), and Cognitive Flexibility (r = .18), all $p <$ .05.

In a follow-up longitudinal study by Kilcullen, Goodwin, Chen, Wisecarver, and Sanders (2002), the TAP predicted the field performance of officers leading their teams through a highly realistic, two-week unconventional warfare field exercise that simulates some of the most difficult activities that operational SF teams are expected to perform. These activities include secretly infiltrating into a 'hostile' area, maneuvering covertly through rugged terrain, executing multiple combat actions and patrols, establishing contact and rapport with guerrilla forces, and arming and training the guerrilla forces to execute multiple combat missions. During the exercise the teams experience levels of calorie and sleep deprivation similar to those encountered in high-intensity operational conditions, and encounter unexpected problems for which there is unclear or incomplete information available to use in developing a solution. As the team leaders, the officers were

ultimately responsible for their team's ability to overcome these challenges and successfully execute the missions. The two criteria consisted of team member and SF cadre ratings of the officers' performance leading his team in the field exercise. Team member ratings of the officer's performance were predicted by the TAP scales of Cognitive Flexibility (r = .39, $p < .05$), Dedication (r = .49, $p < .01$), and Fitness Motivation (r = .42, $p < .01$), a TAP scale developed especially for this population. SF observer ratings were predicted by Dedication (r = .32, $p < .05$) and Fitness Motivation ($r = .43, p < .01$).

Current Selection of Army Special Operations Soldiers

Currently, candidates for three Army special operations units—Rangers, SOF Aviators, and Special Forces—must be selected through a rigorous assessment and selection program prior to joining the unit. Each of the three specialties has unique mission requirements, and therefore has unique requirements for selection. However because there are consistencies across the units in the performance environments in which they operate, there are some common themes in the selection requirements. The typical performance environment for special operations personnel is ambiguous, hostile, and chaotic. Their missions are fast-paced, high risk, and often physically demanding. This requires the selection of soldiers who are able to perform and thrive under these conditions, demonstrating physical strength, endurance, motivation, decisiveness, and a high tolerance for ambiguity and stress. Therefore, one common theme is that candidates are provided limited feedback during the assessment. They are neither encouraged nor discouraged, and are given the opportunity, under moderately high levels of stress, to

demonstrate their physical, emotional, and cognitive capabilities.

75th Ranger Regiment

Candidates for the 75th Ranger Regiment must attend a four-day Ranger Assessment and Selection Program (RASP). The RASP is designed to identify candidates who have sufficient physical capability, motivation, and leadership potential to plan and conduct missions such as direct action raids that seize or destroy identified targets. Candidates complete events such as ruckmarches, an obstacle course, and leadership reaction events, and receive performance ratings from program assessors and peers. A combination of written psychological assessments and interviews by qualified psychologists are used to screen out candidates with psychopathology, unstable or vulnerable personalities, and personal or family situations that would serve as performance distracters.

At the end of the four days, a board of officers and sergeants major from the 75th Ranger Regiment review candidate data and evaluate candidate performance and potential. The focus is on selecting highly motivated, stress tolerant candidates, with the ideal candidate being someone who will persevere in future missions in the face of great personal danger at the extreme limits of their physical and mental abilities. Information about the candidates' potential is presented to the board by Regimental Cadre, the Regiment Psychologist, and Ranger Regimental Officers. In addition, the candidate presents himself to the board and board members can ask questions. The board then votes on each candidate and makes the final selection decision.

Special Operations Aviation Regiment

After serving at least one full tour following the completion of their flight school, candidates for the Special Operations Aviation Regiment attend a one-week assessment program that evaluates their physical and cognitive capabilities, and emotional stability. Cognitive measures include both general intelligence and visual spatial reasoning, which preliminary data show are critical for navigating in flight and reading instrumentation. Assessment exercises evaluate candidates' ability to communicate with the flight crew and to filter emotional reactions while flying. Further, while excessive risk taking would be detrimental to performance, a willingness to take reasonable risks to accomplish the mission is considered beneficial. As with Rangers, psychological assessments and interviews are used to evaluate candidates for psychopathology, unstable personalities, and distracting life stressors, and a board of officers and sergeants major review candidate data. Each candidate's assessment information is presented to the board by senior cadre, SOF Aviators in the Regiment, subject matter experts, and the Regimental Psychologist. In addition, the candidate presents himself to the board and board members can ask questions. The final selection decision is made by the board.

Special Forces

The selection program for the SF is the most lengthy and complex of the three specialties. This is partly due to the multifaceted performance requirements that exist in the SF domain. While Ranger Regiment focuses primarily on direct action (DA) missions, security, and reconnaissance, SF units must be prepared to execute unconventional warfare (UW) and foreign internal defense (FID) missions in addition to DA, reconnaissance, and others. UW and FID missions involve

training and coordinating with either guerrilla or host nation forces in another country. Succeeding in these missions requires interpersonal and cultural adaptability in addition to warrior skills, to ensure that SF soldiers can work effectively in multicultural situations.

The selection for SF is tiered, such that personnel must first meet a number of established prerequisites before attending the multi-week SFAS program. These include established minimum scores for the APFT (229) and the ASVAB GT composite (100), as well as having no record of infractions of the law (e.g., court-martial or Article 15). Research in 1999 documenting the strong predictive relationship between APFT score and success in SFAS led to the increase of the APFT score prerequisite from 209 to the current score required, 229 (e.g., see Zazanis, Hazlett, Kilcullen, & Sanders, 1999).

The SFAS program has evolved since its start in 1988 based both on program commanders' evolving concepts as well as on the results of empirical research (e.g., see Brooks & Zazanis, 1997; Cotty, Bluestein, & Thompson, 2005; Kilcullen, Mael, Goodwin, & Zazanis, 1999; Russell et al, 1995; Sanders, 1996; Teplitzky, 1991; White, Chambers, & Lappin, 2004; Zazanis, Kilcullen, Sanders, & Crocker, 1999; Zazanis, Kilcullen, Sanders, & Litton, 2000).

The SFAS program includes psychological testing and interviews, physically demanding events such as runs and ruckmarches, land navigation exercises, team mission events, and role play exercises. Performance scores from the different events include objective measures such as completion times and number of navigation points found, as well as subjective measures such as assessor and peer ratings. A computer-based analysis system has been developed to assist in scaling the

397scores across the numerous events into performance dimensions. Within this system, assessors enter event scores into handheld computer devices, and MS Access programs scale the data and produce reports.

Similar to other SOF units, SF candidates must demonstrate physical strength and endurance, high cognitive functioning, and leadership qualities such as decision making and influence. In addition, however, SFAS also requires an emphasis on interpersonal skills, cultural sensitivity, and mental adaptability, due to the nature of the SF missions. Finally, these capabilities must be demonstrated under stressful and ambiguous conditions. Table 12.4 lists attributes evaluated during SFAS.

As with the other SOF units, a combination of written psychological assessments and interviews by qualified psychologists are used to screen out candidates with psychopathology, unstable or vulnerable personalities, and personal or family situations that would serve as performance distracters. Written psychological assessments include personality, intelligence, visual spatial ability, abstract reasoning, and achievement tests. After selection, SF candidates still need to complete the rigorous training course prior to joining an SF Group. Attrition analyses have shown that cognitive abilities and achievement tests predict success in the academic phases of the training course, so candidate potential to learn course material is also a key consideration.

Table 12.4. Attributes Screened for During SFAS

Desirable Attributes - Screened In	Undesirable Attributes - Screened Out
Tolerance for stress	Psychopathology
Tolerance for ambiguity	Family vulnerabilities
Situational awareness	Criminal history
Openness	Financial irresponsibility
Cultural sensitivity	Substance abuse
Self-esteem, confidence	Low cognitive ability
Integrity	Inconsistent motivation
Team orientation	Risky lifestyle
Emotional stability	Arrogance/narcissistic
Work motivation	Unethical
Achievement orientation	High need for structure
Fitness motivation	Rigidity
Dominance	Impulsive
High cognitive functioning	Anti-social
Decision-making under stress	Dependency
Adaptive thinking	Anxiety
Interpersonal skills	
Physical fitness	

The selection process differs slightly for officers and enlisted, with additional assessments for officers in the areas of adaptive thinking, situational and other awareness, and self-awareness. At the conclusion of the program, a board of SF officers and sergeants major are presented with candidate data by senior SF cadre, psychologists, and subject matter experts. The board evaluates candidate performance and potential, then makes the final selection decision.

Beyond SF: Transitioning Research Findings to Selection and Classification for the Conventional Army

It is likely that many SOF selection and training practices and procedures would add value to conventional Army units, but the sheer size and composition of the such units makes such a transition difficult. One exception has been the use of the TAP in settings other than SF (see Table 12.5 for details). The promising results obtained with the TAP and its follow-on test, the Rational Biodata Inventory (RBI), have led to operational implementation in some settings. For example:

- The efforts with Army civilian supervisors led to the incorporation of several TAP scales into the Civilian Leader Improvement Battery, a web-based assessment and self-development system available to all Department of the Army and Department of the Navy civilian employees.
- As a result of work done at the Army War College, TAP scales were implemented for use there and at the Industrial College of the Armed Forces as part of an executive assessment battery that provides self-development feedback to students at these institutions.
- Research with Army enlisted soldiers that included both TAP and the RBI led to the incorporation of these scales into a new assessment battery intended to supplement the ASVAB in the enlisted accession process. The TAP/RBI scales are also part of a battery currently used at one reception battalion to select soldiers for accelerated BCT.

Table 12.5. TAP Validation Studies in Conventional Army Forces

Study	Sample	Criteria	Validation Results
Kilcullen (2006)	1664 Department of Army Supervisors	Confidential off-the-record performance ratings by supervisors Administrative data (e.g. bonuses, letters of commendation, pay raises, disciplinary actions)	Supervisory ratings of leader performance predicted by ST (r = .13**), D, (r = .15**), PL, r = .20** Administrative indices predicted by CF (r = .15**), D (r = .25**), TP (r = .18**), PL (r = .23**)
Kilcullen, White, Zaccaro & Parker (2000)	340 civilian supervisors, managers and Senior Executive Service (SES) leaders	Supervisory ratings of job performance	Supervisory performance predicted by D (r = .32**), PL (r = .26**) Managerial performance PL (r = .23**), CF (r = .20**), D (r = .19**) SES performance PL (r = .29**)
Kilcullen, Putka, & McCloy (2007)	688 enlisted soldiers in their first term	Supervisor ratings of technical performance, AE, PT, T, expected future performance	PL (median r = .19), SE (median r = .19), PD (median r = .18), FM (median r = .17), AO (median r = .15)
Putka & Le (2005)	2,640 enlisted soldiers	Attrition from Initial Entry Training	FM (r = -.20*), SE (r = -.10*), AO (r = -.07*)

Note: ST = Stress Tolerance; D = Dedication; PL = Peer Leadership; CF = Cognitive Flexibility; TP = Team Player; SE = Self Esteem/Self Efficacy; PD = Personal Discipline; FM=Fitness Motivation; AO = Achievement Orientation; AC = Army Affective Commitment; CT = Cultural Tolerance; AE = Achievement/Effort; PT = Physical Fitness; T = Teamwork
** $p < 05$, ** $p < .01$*

Table 12.5. TAP Validation Studies in Conventional Army Forces (Continued)

Study	Sample	Criteria	Validation Results
Hoffman & Muraca (2007)	152 BCT soldiers	Army commitment	SE (t = 5.27*), PD (t = 2.68*) uniquely predicted commitment beyond GT score
Heffner (2007)	123 soldiers in accelerated BCT	Drill sergeant evaluations of overall effectiveness	Multiple R = .46* obtained from FM, SE, AO, ST, AC and CT plus ASVAB GT score and APFT score
Klopp (2006)	647 college juniors in ROTC Leader Training Course	ROTC completion	AC (r = .22*), ST (r = .15*), PL (r = 09*), ID (r = 09*)
Kilcullen, White, Sanders, & Hazlett (2003)	298 MOS 95C correctional specialists	Supervisor ratings of job performance Prison rules violations Supervisory ratings of willingness to stay out of trouble and to follow rules and regulations	Job performance predicted by RA (r = .30*) and SM (r = .14*) Rules violations predicted by SM (r = .22*) and RA (r = .25*) Supervisory ratings of staying out of trouble predicted by SM (r = .20*) RA (r = .30*)

Note: ST = Stress Tolerance; D = Dedication; PL = Peer Leadership; CF = Cognitive Flexibility; TP = Team Player; SE = Self Esteem/Self Efficacy; PD = Personal Discipline; FM=Fitness Motivation; AO = Achievement Orientation; AC = Army Affective Commitment; CT = Cultural Tolerance; AE = Achievement/Effort; PT = Physical Fitness; T = Teamwork
** p < 05, ** p < .01*

Table 12.5. TAP Validation Studies in Conventional Army Forces (Continued)

Study	Sample	Criteria	Validation Results
Kilcullen, White, Sanders & Hazlett (2003)	252 prisoner soldiers	Disciplinary incidents while in jail Number of parole violations	Disciplinary incidents predicted by SM (r = -.36**), RA (r = -.27*) Parole violations predicted by SM (r = -.14*) RA (r = -.12*)
Kilcullen, White, Sanders & Hazlett (2003)	495 Army soldiers entering IET	Responses to questions about pre-service troubles (e.g. prior theft, drug use, criminal activity)	Drug use, criminal activity, and job-related disciplinary problems predicted by SM (median r = -.27) RA (median r = .18)

Note: SM = Social Maturity; RA = Respect for Authority
** p < 05, ** p < .01*

Summary

From their early roots in the OSS, it was clear to Army personnel involved in SOF activities that standard selection measures would not be sufficient to identify personnel with the requisite physical and psychological traits to function in the challenging environments they would likely confront. Therefore, throughout its history, SOF selection has involved a much broader range of testing that befits the special nature of these units. Investing time and energy into the development and implementation of effective selection methods for these units is essential, given the strategic importance of their missions. To the extent that products of these efforts can be successfully applied to selection needs in conventional forces, both groups ultimately benefit.

References

Abelson, H. I. (1954). *Factors related to the effectiveness of Special Forces personnel.* Washington, DC: George Washington University.

Bank, A. (1986). *From OSS to Green Berets.* Novato, CA: Presidio Press.

Banks, L. M. (1995). *The Office of Strategic Services psychological selection program.* Ft. Leavenworth, KS: Command & General Staff College.

Berkhouse, R. G. (1963). *Research on Combat Selection and Special Forces Manpower Problems – Status Report* (AFRC Research Study No. 63-2). Washington, DC: U.S. Army Personnel Research Office.

Berkhouse, R. G., & Cook, K. G. (1961). *Development of preliminary screening measures for Special Forces trainees* (Research Memorandum No. 61-7). Arlington, VA: Army Behavioral Evaluation Laboratory.

Berkhouse, R. G., Mendelson, M. A., & Cook, K. G. (1961). *Development of performance measures of individual proficiency in Special Forces* (Research Memorandum No. 61-6). Arlington, VA: Army Behavioral Evaluation Laboratory.

Brooks, J. E., & Zazanis, M. M. (1997). *Enhancing U.S. Army Special Forces: Research and Applications* (Special Rep. No. 33). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Burke, W. P., & Dyer, F. N. (1984). *Physical fitness predictors of success and injury in Ranger training.* Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Clifton, T. C., Kilcullen, R. N., Reiter-Palmon, R., & Mumford, M. D. (1992). *Use of a triple cross-validation technique on alternative biodata scaling procedures.* Paper presented at the 100th annual American Psychological Association conference, Washington, DC.

Cotty, W., Bluestein, B., & Thompson, J. (2005). The whole-man concept: Assessing the SF soldier of the future. *Special Warfare, 17,* 18-21.

George, E. L, & Cassady, P. D. (1981). *Training Attrition Problem, Institute for Military Assistance Training Subsystem Effectiveness Analysis* (TRASANA TFA-13-81). White Sands Missile Range, NM: U. S. Army TRADOC Systems Analysis Activity.

Heffner, T. (2007). *ARI Selection Research for BCT Multi-Tracking.* Briefing presented to the U.S. Army Training Center, Fort Jackson, SC.

Hoffman, R. R., & Muraca, S. T. (2007). *Cognitive and Non-cognitive predictors of Basic Combat Training performance and attitudes.* Paper presented at the 115th annual American Psychological Association conference, San Francisco, CA.

Kilcullen, R. N. (2006). *Enhancing DoD civilian leader development with the Civilian Leader Improvement Battery (CLIMB).* Paper presented at the 2006 Army Science Conference, Orlando, FL.

Kilcullen, R. N., Goodwin, J, Chen, G., Wisecarver, M., & Sanders, M. (2002). *Identifying agile and versatile officers to serve in the objective force*. Paper presented at the 2002 Army Science Conference, Orlando, FL.

Kilcullen, R. N., Mael, F. A., Goodwin, G. F., & Zazanis, M. M. (1999). Predicting U.S. Army Special Forces field performance. *Human Performance in Extreme Environments, 4*, 53-63.

Kilcullen, R. N., Putka, D. J., & McCloy, R. A. (2007). Validation of the Rational Biodata Inventory (RBI) for predicting the job performance of U.S. Army enlisted soldiers. In D. J. Knapp. & T. R. Tremble (Eds.), *Concurrent validation of experimental Army enlisted personnel selection and classification measures* (Tech. Rep. No. 1205, pp. 117-134). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA 471963)

Kilcullen, R. N., Putka, D. J., McCloy, R. A., & Van Iddekinge, C. H. (2005). Development of the Rational Biodata Inventory (RBI). In D. J. Knapp, C. E. Sager, & T. R. Tremble (Eds.), *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria* (Tech. Rep. No. 1168, pp. 105-116). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Kilcullen, R. N., White, L. A., Mumford, M. D., & Mack, H. (1995). Assessing the construct validity of rational biodata items. *Military Psychology, 7*, 17-28.

Kilcullen, R. N., White, L. A., Sanders, M., & Hazlett, G. (2003). *The Assessment of Right Conduct (ARC) administrator's manual* (Research Note No. 2003-09). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Kilcullen, R. N., White, L. A., Zaccaro, S., & Parker, C. (2000). *Predicting Managerial and Executive Performance*. Paper presented at the 15th annual Society for Industrial and Organizational Psychology conference, New Orleans, LA.

Klopp, R. G. (2006). *Predicting commissions of attendees of the 2004 U.S. Army Reserve Officer Training Corps (ROTC) Leader's Training Course (LTC)*. Gaithersburg, MD: Advanced Engineering & Planning Corporation.

Lock, J. D. (1998). *To fight with intrepidity: The complete history of the U.S. Army Rangers 1622 to present*. New York, NY: Simon & Schuster.

MacKinnon, D. W. (1980). *How assessment centers were started in the United States: The OSS Assessment Program*. Pittsburgh, PA: Development Dimensions International.

Marder, M., & Medland, F. F. (1964). *Construction of experimental aptitude battery for Special Forces enlisted personnel* (Research Memorandum No. 64-9). Washington, DC: U.S. Army Personnel Research Office.

Marder, M., & Medland, F. F. (1965). *Analysis of experimental selection tests for Special Forces officer training* (APRO Research Study No. 65-3). Arlington, VA: Army Behavioral Evaluation Laboratory.

Morgan, W. J. (1957). *The O.S.S. and I.* New York, NY: Norton.

Morgan, W. J. (1994). Personal communication to the author.

Murray, H. A., & MacKinnon, D. W. (1946). Assessment of OSS personnel. *Journal of Consulting Psychology, 10,* 76-80.

Olmstead, J.A., Caviness, J. A., Powers, T. R., Maxey, J. L., & Cleary, F. K. (1972). *Selection and training for small independent action forces: Final report* (TR-72-2). Alexandria, VA: Human Resources Research Organization.

Office of Strategic Services Staff (1948). *The assessment of men.* New York, NY: Rinehart.

Pleban, R. J., Allentoff, H. L., & Thompson, T. J. (1989). *Preliminary assessment of selected predictors of Special Forces Qualification Course success* (Research Report No. 1539). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Pleban, R. J., Thompson, T. J., Valentine, P. J., Dewey, G. I., Allentoff, H. L., & Weslowski, M. A. (1988). *Selection and Assessment of Special Forces Qualification Course candidates: Preliminary issues* (Research Note No. 88-37). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Putka, D. J. & Le, H. (2005). *Select21 Attrition update: Initial entry training attrition* (Research Rep.). Alexandria, VA: Human Resources Research Organization.

Roosevelt, K. (1976). *War report of the OSS.* New York, NY: Walker and Co.

Russell, T. L., Crafts, J. L., Tagliareni, F. A., McCloy, R. A., & Barkley, P. (1996). *Job analysis of Special Forces jobs* (Research Note No. 96-76). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA324584)

Russell, T. L., Rohrback, M. R., Nee, M. T., Crafts, J. L., Peterson, N. G., & Mael, F. (1995). *Development of a roadmap for Special Forces selection and classification.* (Tech. Rep. No. 1033). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA317151).

Sanders, M. (1996, Fall). Special Forces Assessment and Selection attrition analyses. *ARI Newsletter, 6* (3), 13-14.

Schoenfeldt, L. F. (1989). *Biographical data as the new frontier in employee selection research.* Address presented at the annual meeting of the Division of Evaluation, Measurement and Statistics of the American Psychological Association, New Orleans, LA.

Simpson, C. M. (1983). *Inside the Green Berets: The first thirty years.* Novato, CA: Presidio Press.

Teplitzky, M. L. (1991). *Physical performance predictors of success in Special Forces Assessment and Selection* (Research Rep. No. 1606). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

The OSS Assessment Staff, *Assessment of Men: Selection of Personnel for the Office of Strategic Services* (New York: Rinehart & Co., 1948; reprint, New York, NY: Johnson Reprint Corp., 1978), 13.

Thompson, T. J., Pleban, R. J., Valentine, P. J., Dewey, G. I., Wesolowski, M. A., & Allentoff, H. L. (1987). *Special Forces Qualification Course: Exploratory research for selection and development.* Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Uhlman, C. E., Reiter-Palmon, R., & Connelly, M. S. (1990). *A comparison and integration of empirical keying and rational scaling of biographical data items.* Paper presented at the 36th Annual Southeastern Psychological Association conference, Atlanta, GA.

White, L. A., & Kilcullen, R. N. (1992). *The validity of rational biodata scales.* Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

White, S. S., Chambers, B. A., & Lappin, M. (2004). *Special Forces manpower fill: Training pipeline attrition statistics and recommendations.* (Final Contractor Rep. No. 462). Arlington, VA: Personnel Decisions Research Institutes, Inc.

Zazanis, M. M., Hazlett, G. A., Kilcullen, R. N., & Sanders, M. G. (1999). *Prescreening Methods for Special Forces Assessment and Selection.* (Tech. Rep. No. 1094). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Zazanis, M. M., Kilcullen, R. N., Sanders, M. G., & Crocker, D. A. (1999, Summer). Special Forces selection and training: Meeting the needs of the force in 2020. *Special Warfare, 12*, 22-31.

Zazanis, M. M., Kilcullen, R. N., Sanders, M. G., & Litton, D. A. (2000, Fall). The SF pipeline review: Voices from the field. *Special Warfare, 13*, 6-16.

CHAPTER 13

THE SELECTION AND CLASSIFICATION OF ARMY AVIATORS

Teresa L. Russell
Human Resources Research Organization (HumRRO)

Michael G. Rumsey
U.S. Army Research Institute for the Behavioral and Social Sciences

World War I

In April of 1917, when the US entered World War I (WWI), the Aviation section of the Army Signal Corps had 52 trained flyers (Henmon, 1919). By the end of the war, 1,200 cadets began training every month (Thorndike, 1919). The problem of selecting and classifying aviators was not just one of volume; little was known about the abilities required of aviators. There was great difficulty in "prophesying abilities in a school [aviator] such as had never before existed, in an art which only a few score men in the country had learned, and in a form of warfare which was only three years old and was changing its nature radically every few months" (Thorndike, 1919, p. 1255). To meet the needs of the war, the National Research Council (NRC) organized the Psychology Committee and 12 subcommittees, one of which was established to study the psychological problems of aviation, including examination of aviation recruits (Yerkes, 1919).

The NRC's aviation subcommittee was charged with aiding the Air Service in the selection, training, and maintenance of pilots. The Air Service created the Examining Boards and Physical Examining Units and established physical, moral, and medical standards for pilots, observers, and balloonists. They adopted *Thorndike's Mental Alertness Test* (a test of general mental ability) as a regular part of the examination procedure. There was much concern about the desired qualities of pilots. Henmon (1919) reported that he and other staff were repeatedly told that the "flying officer was not to be an aerial chauffeur but a twentieth century cavalry officer mounted on Pegasus" (p. 103).

Selected trainees were sent to ground school (a theoretical and practical course on military aeronautics) and then to flying school. "Examiners soon found that while it was relatively easy to pick men who could negotiate the non-Pegasarian flights demanded in the ground schools, what they would do when they mounted Pegasus was a sheer guess" (Henmon, 1919, p. 103-104). Wash-out rates were high and costly. One trainee wrecked five planes within 55 hours of flying instruction before being labeled "unadaptable for further training."

Thorndike's research indicated that scores on the *Thorndike's Mental Alertness Test* correlated .50 with performance in ground school but were less predictive of performance in flying school (E.L. Thorndike, 1919). To better predict aptitude for flying, the aviation subcommittee initiated a series research projects to develop mental and physiological tests. They compiled an initial battery of tests based on the work of French and Italian psychologists as well as some of their own thinking. Different combinations of test batteries were tried out in sequential studies on (a) 75 cadets in Ground School at the Massachusetts Institute of Technology, (b) 50 flying cadets at

San Diego, and (c) 44 cadets at Essington Field in Pennsylvania. Refinements to the test battery were made with each try-out. Based on the results, 10 tests (listed in Table 13.1) were selected and given to a total of 300 cadets and flying instructors at Kelly Field, San Antonio and Rockwell Field, San Diego. The 150 participants at each field were pre-selected as follows: 50 very good flyers, 50 very poor flyers, and 50 with unknown flying skills. Officers in Charge of Flying rated the participants' performance based on progress cards and judgments of stage commanders.

For unstated reasons, *Thorndike's Mental Alertness Test* and the *Athletic Achievement and Interest* form were not administered to a sufficient number of participants at Rockwell Field. But they were both recommended and adopted for use based on other data. The *Athletic Achievement and Interest* score correlated more than .60 in the Kelly sample, with less than 100 cases, and was thought to be suspiciously high given that correlations for other measures were much lower.

Today, we would probably correct the correlations between test scores and performance for direct and indirect range restriction on the *Thorndike's Mental Alertness Test* since it was in use operationally. Regardless, *Thorndike's Mental Alertness Test* and the *Test of Emotional Stability* (in particular the amplitude of hand movement in response to a pistol shot) yielded the highest correlations with flying ability. *Perception of Tilt* and *Swaying* were also among the top predictors. A composite of scores on selected tests yielded a multiple correlation coefficient of .70. The author noted that the research team would have preferred to create a composite using regression weights, but time and labor for that analysis were not available (Henmon, 1919, p. 107). With no computers, computation of

Table 13.1. Ten Tests Administered at Kelly and Rockwell Fields in 1918

Name	Definition
Emotional Stability	Measured (a) the amplitude of hand tremor, (b) pulse rate, (c) amplitude and frequency of respiration, and (d) arithmetic performance in response to the unexpected discharge of a pistol.
Perception of Tilt	Measured sensitivity to gradual changes of body position using a specially designed tilting chair.
Swaying	Measured motor coordination and control while the subject was required to stand steadily with his eyes open and in turn with his eyes closed.
Visual Reaction	Measured quickness of response to the movement of the upper hand on the dial of a chronoscope.
Auditory Reaction	Measured quickness of response to the noise of a sound hammer.
Equilibrium Reaction	Measured quickness of response to the noise of sudden changes in body position as the subject stood on a tilting table.
Equilibrium Differential	Measured the relative quickness of the response to changes of equilibrium compared to simple reaction times by taking the sum of the visual and auditory reaction times minus the equilibrium reaction time.
Extension of Curves	Measured the ability to continue, in imagination, certain segments of curves presented visually. It was intended to simulate the judgments of distance required in landing a plane.
Thorndike's Mental Alertness Test	13 tests of various functions designed to measure general mental ability. This test was a part of the selection procedure.
Athletic Achievement and Interest	An application blank that asked biographical and interest questions, in particular about athletics.

Note. Adapted from Henmon (1919).

correlation matrices and advanced analyses were all done by hand. As it turned out, the tests were not fully implemented before the armistice was signed and schools were closed.

Members of the committee at the Medical Research Laboratory at Hazelhurst Field conducted the first studies of aviation medicine. They developed a series of tests to assist in determining the ability of aviator candidates for withstanding the effects of high altitudes. The general approach called for the subject to perform attention and motor tasks while his oxygen supply gradually decreased. Performance decrements were noted up to the moment of "complete inefficiency" which would have resulted in unconsciousness had the subject not received air. These researchers also experimented with apparatus, or psychomotor, tests.

World War II[114]

After WWI, the need for military aviators waned, but civilians were becoming more interested in flying. In 1923, the Army had only 17 applicants for aviation cadet training, 5 of whom were selected. Applications increased progressively during the '20s to 363 in 1928, 5 of whom were selected (Davis, 1947). During this era, the Army could be selective, and the criteria for selection were fairly stringent. Men could qualify for aviation cadet training only if they had completed at least two years of college, or had passed a special written examination covering nine basic college subjects. These examinations began in 1927 and remained virtually unchanged until 1941. In the spring of that year, there was a great expansion in the training program

[114] Virtually all of the material in this section is based on The Army Air Forces Aviation Psychology Program Research Reports, an incredibly rich and informative 19-volume series of reports. Most of the material is from Report 1 (Flanagan, 1948), Report 2 (DuBois, 1947), Report 3 (Thorndike, 1947), and Report 6 (Davis, 1947).

for Army Air Force (AAF) pilots and other air-crew specialties, and it became apparent that the increasing demand could not be met with the existing selection requirements.

The Aviation Psychology Program in the AAF began in July of 1941 when the Medical Division of the Air Corps commissioned Dr. John Flanagan as a major to direct the psychological research unit. The program grew to include hundreds of men and women, many of whom were, or would become well-known aviation, industrial, or educational psychologists. Some of the most prominent AAF aviation psychology researchers were Edwin Ghiselli, Lloyd Humphreys, J.P. Guilford, Robert M. Gagne, Paul Horst, Arthur Melton, John Lacey, Robert L. Thorndike, S. Rains Wallace, and Meredith Crawford, to name a few. Julius E. Uhlaner, who would later become Chief Psychologist of the U.S. Army and the first Technical Director of the Army Personnel Office, was an enlisted man on the team.[115]

By November 1941, it was clear that the Army could not meet the demand for air crew personnel without drastic changes to the selection system. Eventually 563,916 aviator cadet applicants would be tested in 1942, all of whom were volunteers; 293,588 were selected for training. The AAF needed to replace the education requirement with an exam that could be administered (a) to groups of applicants in a reasonable amount of time, (b) with little in the way of apparatus or special materials, (c) by individuals with no professional training in psychology, and (d) in hundreds of Aviation Cadet Examining Boards scattered throughout the U.S. and in Army bases overseas. Eventually, it

[115] DuBois (1947) provides a list of hundreds of officers, enlisted men, women in the Women's Army Corp (WACs), and on-the-line trainees who were involved in test development, research, or administration.

also became apparent that testing was needed, for both selection of air-crew for training and for classification into a specialty—pilot, navigator, or bombardier.

The Aviation Psychology Program developed a two stage-process for selection and classification. First, candidates took the *AAF Qualifying Examination*[116], a paper-and-pencil power test that took about three hours to administer, and went through rigid medical screening. After passing this initial selection, cadets were admitted to aviation-student status and sent to a classification center to take the *Aircrew Classification Battery*. Scores on this test yielded composite aptitude scores designed to predict graduation or elimination from pilot, bombardier, or navigator flying training.

The AAF Qualifying Examination

The Aviation Psychology Program held a conference of experts in test construction on 18-19 December 1941 to create the final outline for the *Qualifying Examination* and to review items drafted by staff. Immediately after the conference the first form of test (AC10A) was printed and put into official use on 15 January 1942. The machinery for ongoing research and test maintenance was put in place. Staff routinely conducted psychometric analyses of items and when training data were available, correlated items with pass/fail training performance. By the end of the war, 17 forms containing 2,910 different test items had been created.

The outline for the first version of the test (AC10A) is shown in Table 13.2. The next seven forms (AC10B-AC10H) were only

[116] The *AAF Qualifying Examination* was first published as the *Aviation Cadet Qualifying Examination* on 15 January 1942. It was renamed on 1 June 1944 because its use was extended to the selection of enlisted men to serve as gunners in the aircrew.

minor modifications of the first one. It was a highly reliable test; the reliability of the total score was .93 in a sample of 370 applicants, with reliabilities of subtest scores ranging from .36 for Practical Judgment to .88 for both Vocabulary and Mathematics. The cut point was set on the total test score, not subtest scores.

Table 13.2. Original Outline for the Qualifying Examination

Test	Number of Items
Verbal Tests:	
General vocabulary	45
Reading comprehension	15
Information Tests:	
Contemporary affairs in aviation and the war	30
Miscellaneous Tests:	
Practical judgment	15
Mathematics	30
Mechanical comprehension	15

Note. Adapted from Davis (1947).

Based on data gathered in training, fairly drastic changes were made in the next three forms (AC12I-AC12K). Mathematics items were dropped because there was a greater need for pilots than navigators (for whom mathematics was most relevant), and mathematics items were included on the classification batteries. In short, these forms contained no verbal or mathematics tests, doubled the number of Mechanical Comprehension items over previous editions, and included four perceptual tests—Planning

Circuits, Hidden Figures, Point Distance, and Path Distance. The overall length of the test went from 150 to 270 items. Prior forms of the test had relied more heavily on power than on speed; applicants simply took the test at their own pace and turned it in when they were done. In the AC12 series forms, the perceptual tests were speeded and had time limits.

One other overhaul occurred for the remaining six forms. Reading Comprehension was added back in, all of the perceptual tests except Hidden Figures were dropped, and short tests of flying and driving information were added. Research suggested that the Hidden Figures test was about as predictive of graduation with or without time limits, so the requirement for a time limit was dropped. This reduced the administrative burden of separately timing tests. The total number of items dropped back to 150.

Aircrew Classification Battery

Classification is a more complex problem than selection. The goal is to maximize the overall effectiveness of the organization by assigning people to the jobs that they are most likely to perform well, within constraints on available training seats. If assignments had been left to individual preferences, bombardier and navigator seats would have been difficult to fill. Most men wanted to become pilots. Initially, the AAF filled navigator and bombardier training with individuals who had failed pilot training. The purpose of the *Aircrew Classification Battery* was to allow selection of navigators and bombardiers before sending men to pilot training so as to maximize the efficiency of the assignment decision. That is, the goal was to maximize the average predicted performance of cadets for each aircrew assignment.

To accomplish this goal, the tests in the battery needed to differentiate the aircrew assignments. Based on job analyses of pilot, bombardier, and navigator jobs, classification researchers developed lists of constructs to be measured and organized them according to different fields and areas within those fields, as shown in Table 13.3.

Table 13.3. Major Fields and Areas of Content for Aircrew Classification Batteries

Major Fields	Areas within Fields
Tests of emotion, temperament, and personality	Absence of tenseness, Absence of confusion and nervousness, Absence of fear and apprehension, Motivation, Personal information, Projective techniques, Fatigue
Tests of intellectual functions	Information, Reasoning, Judgment, Foresight and planning, Memory, Comprehension, Mathematics, Physics, Mechanical intelligence.
Tests of motor performance	Gross coordination, Fine muscle coordination, Appropriateness of controls used, Feel of Controls, Smoothness of control movement, Progress in developing technique, Serial learning, Pursuit.
Perceptual tests	Visualization of flight course, Estimation of speed and distance, Sense of sustentation, Division of attention, Orientation, Speed of decision and reaction, Auditory discrimination, Form perception.
Physical measures and sensory tests	Size, Weight, Night vision, Depth vision, Color vision.

Over the course of the war, AAF researchers implemented 10 *Classification Batteries*. Content differences across batteries are summarized in Table 13.4. Initially, the battery was composed

of available tests that were expected, based on job analysis results, to differentiate aircrew jobs. Initial batteries had a fairly large number of perceptual tests, most of which were speeded in the early days. As shown, the number of perceptual tests gradually decreased as information about their intercorrelations and correlations with criterion data became available. The other major change had to do with the addition of measures of emotion, temperament, and personality traits. In the August 1942 battery, these constructs were measured with information and apparatus tests (e.g., Steadiness Under Pressure which was included despite differences of opinion among directors about its value). But, later batteries included biographical history items to tap these constructs.

Table 13.4. Numbers of Tests from Different Fields in Aircrew Classification Batteries

Classification Battery	Tests of emotion, temperament, and personality	Tests of intellectual functions	Tests of motor performance	Perceptual tests
February 1942	0	5	4	13
April 1942	0	5	4	9
May 1942	0	5	4	9
June 1942	0	5	4	8
August 1942	4	6	3	6
December 1942	4	6	4	6
July 1943	5	5	3	6
November 1943	4	8	5	5
September 1944	3	8	4	6
June 1945	3	8	5	7

The tests were also differentiated by their mode of administration. Of the 18 to 22 tests in each battery, 4 or 5 were apparatus tests; the rest were printed instruments. While most apparatus tests were used to measure motor performance (e.g., Rudder Control), some measured perceptual ability (e.g., SAM Rotary Pursuit with Divided Attention) or even emotion, temperament, and personality (e.g., Aiming Stress). Apparatus testing added a great deal of complexity to the testing system. It required elaborate and careful standardization of equipment and procedures and close supervision by examiners. Four candidates could take any one apparatus test at a time (each on a separate apparatus) with one examiner present. Each apparatus had to be calibrated and maintained. Equipment specifications were very important; Melton (1947) provides the schematic wiring diagrams for several of the tests. Over the years, it took time to procure and install new apparatus tests. Therefore, in later years of the war, more and better apparatus tests were available.

The Thousand Applicants Study

Despite a good deal of resistance from operational units, in May 1943 Aviation Psychology Program researchers convinced the Chief of the Psychology Branch to examine a large sample of applicants with both the *Qualifying Examination* and the *Classification Battery* and to send **all** of the physically qualified applicants tested into training regardless of the test results. The study was designed to yield at least 1,000 participants; in the end, there were 1,143 valid cases in the sample. All of these men, sampled from different sections of the country, took the *Qualifying Examination*, went to basic training, were sent to a classification center for classification testing, and were sent into pilot preflight school regardless of their test scores. They were sent to training along with other cadets who were not in the study with no

designation as to which ones were members of the experimental group.

The results provided strong support for the test batteries. The biserial correlation between the stanine score on the *Qualifying Examination* and graduation or elimination from training was .64. The two printed tests that were the best predictors of graduation were the General Information Test and Instrument Comprehension Test II, with biserial *r*s of .51 and .48, respectively.

Over the course of the many forms of the *Qualifying Examination*, researchers had tried out several information tests. Information tests were knowledge tests that were thought to be surrogate measures of interest. "A man tends to learn more and to remember more about things that interest him than about things to which he is indifferent" (Davis, 1947, p.103). The General Information test asked questions about technical information that would be acquired by applicants having interests relevant to air-crew positions (e.g., The plane with a cannon in its nose is manufactured by: (a) Bell (b) Boeing (c) Sikorsky, (d) Douglas, (e) Vultee). Instrument Comprehension II was a spatial test that presented drawings of two instruments, a compass and an artificial horizon, and asked the examinee to choose which of five pictures depicted a plane positioned according to the instrument readings. Other printed tests with predictive value were the Mechanical Principles Test, Spatial Orientation Test II, the Biographical Data Blank (Pilot form), and Spatial Orientation Test I.

Six apparatus tests were in use at the time of the study. With exception of the Finger Dexterity Test, all of them provided useful validities for predicting the graduation criterion—the Discrimination Reaction Time Test (r_{bis} = .42), Rudder Control

Test (r_{bis} = .40), Complex Coordination Test (r_{bis} = .41), Two-Hand Coordination (r_{bis} = .36), and Rotary Pursuit (r_{bis} = .31).

Scientific Contributions

By the end of the war, the AAF Aviation Psychology Program included about 200 officers, 750 enlisted men, and 500 civilians. Together, this group made significant contributions to the fields of industrial and military psychology. This section has only touched on the main studies accomplished during the war. Several of the key AAF Aviation Psychology contributions to the science of employee selection and classification are listed briefly below:

- Correlational and factorial research suggested that the number of measurable, distinct traits was much higher than originally thought.
- Classification analyses suggested that differential patterns of aptitude requirements existed for different types of activities and that those patterns were useful for placement of individuals into jobs.
- Objective tests were found to have greater predictive power than scores on interviews, projective methods, or ratings.
- Multiple-choice tests with separate answer sheets were highly economical and effective testing tools.
- Objective, multiple-choice responses could be gathered for tests presented by way of motion pictures.
- Calibration and maintenance procedures were developed to ensure the standardization and accuracy of apparatus tests.

- Studies of subjective clinical procedures for obtaining information about individuals indicated that few of these procedures had validity for predicting success in air-crew specialties.
- There were advantages to untimed or power tests. They were less burdensome to administer than timed tests and provided greater predictive power.

Finally, the group made significant contributions to techniques of prediction and experimentation, the theory and knowledge of the design of an apparatus, and the theory and knowledge of education and training.

Post-World War II

Flight Aptitude Selection Test

In 1947, the Air Force became a separate service, equivalent with the Army and Navy. With it went most of the fixed-wing aircraft that had belonged to the Army Air Corps. However, helicopters, or rotary wing aircraft, continued to play a significant role in the Army, which retained some fixed wing aircraft as well. Most rotary wing pilots were warrant officers, while fixed wing pilots tended to be commissioned officers. Research activity to improve selection of Army pilots began in 1955 (Drucker & Kaplan, 1966), just before training of Army pilots was passed from the Air Force to the Army in 1956 (Wesolek, 2007).

Selection of fixed wing pilots did not present a major challenge. In the late 1950s, a test battery based on existing Air Force tests, known as the AFWAB-1, was administered to Army students and found to be an effective predictor of success in multiple flight training contexts. "It included background inventory, aeronautical information, mechanical principles,

aircraft orientation, and flight visualization" (Drucker & Kaplan, 1966, p. 30).

Although more attention was given to selection of rotary wing pilots, particularly because early experience indicated that many warrant officers had difficulty demonstrating the level of leadership effectiveness needed by pilots, here again there was initially heavy reliance on research conducted by the other services. A combination of existing tests and newly developed tests, 40 in all, were examined in a long-term testing program. As the program proceeded, interim findings were adapted to immediate needs. "Between 1965 and 1961, three interim batteries to help select helicopter pilot trainees were successively implemented (Drucker & Kaplan, 1966, p. 31)." The third of these, the ARWAB-1, consisted of a number of measures, including "the locations test, the complex movements test, the helicopter pilot self-description form, and the helicopter information test (Drucker & Kaplan, 1966, p. 31)." It was determined to be sufficiently effective that it could be used until the long term research effort was completed.

The culmination of the research program was the implementation of the Flight Aptitude Selection Test (FAST) in 1966. The FAST consisted of 12 tests, which could be combined into one battery for use in selecting commissioned officers, primarily as fixed wing aviators and another battery for use in selecting warrant officers as rotary wing aviators. Eight tests were shared by both batteries, while two tests, Mechanical Principles and Flight Orientation, appeared only in the officer battery. Each battery contained a non-cognitive self report measure as well—a Biographical Information test for officers, and a Self-Description measure for warrant officers. The twelve tests could be group into "four basic content areas: 1)

biographical data and interest information, 2) spatial ability, 3) mechanical ability, and 4) aviation information" (Brown, Dohme, & Sanders, 1982, p. 1174). Research conducted on students in Initial Entry Rotary Wing training in 1974 revealed a .38 validity for the warrant officer battery and a .44 validity for the commissioned officer battery against a training success outcome measure (Eastman& McMullen, 1978). An updated and shorter version (Revised FAST (RFAST)) was implemented in 1980, with a validity of .33 against training grades. This revised version, which was the same for both commissioned officers and warrant officers, had only 7 tests (Brown, Dohme, & Sanders, 1982). In 1988, a new version, the Alternate Flight Aptitude Selection Test (AFAST; Katz, 2006), consisted of the following components: Background Information Form, Instrument Comprehension Test, Complex Movements Test, Helicopter Knowledge Test, Cyclic Orientation Test, Mechanical Functions Test, and Self-Description Form (Department of the Army, 1987). The AFAST was validated in 1993 (Cross, 1997).

Classification Battery

Helicopter training proceeds through a series of stages. Prior to 1988, the first stage involved use of a light training helicopter, the TH-55 Osage. The second stage involved training on UH-1 Iroquois, a utility helicopter popularly known as the "Huey," which was featured prominently in the Vietnam War. Alternatively, a small number of students received training on the OH-58 Kiowa, a small scout helicopter. Assignment to, and training on, any other operational Army helicopter did not occur until after flight school was completed.

In 1988, a change in the manner in which helicopter training was conducted gave rise to the need for a rotary wing

classification battery. That year, the Army Aviation Center initiated the Initial Entry Rotary Wing Multi-Track program of instruction. Under this program, training on the UH-1 Huey became the first stage, replacing TH-55 Osage training. Prior to completion of this stage, students were assigned to one of four tracks for the remainder of their training: 1) the UH-1; 2) the AH-1, an attack helicopter; 3) the OH-58 Kiowa, or 4) the UH-60 Blackhawk, a utility helicopter which had begun to replace the UH-1 in Army operational use in 1979 (Intano, Howse, & Lofaro, 1991a).

With differences in design, operating characteristics, and missions of the various helicopters, there was a reasonable expectation that the individual skills, knowledge and abilities associated with success in the operation of one type of helicopter might be different than those associated with success in another. Job analyses were conducted to identify tasks and personal traits associated with each helicopter type, as well as those common across type. Existing test batteries "were considered and evaluated for their potential to discriminate among aviators already highly qualified in different aircraft" (Intano, Howse, & Lofaro, 1991b, p. 1). Due to time pressures associated with the need to quickly implement the new classification battery, these activities were conducted concurrently.

Tests were drawn from the Air Force Basic Abilities Test, a Navy battery, a National Aeronautics and Space Administration battery, and an Army battery developed for another purpose. The resulting battery was a combination of heterogeneous tests assessing cognitive abilities (e.g., verbal, math, information, reasoning, pattern recognition and detection), personal traits (e.g., composure/stress handling, decisiveness,

adaptability/flexibility), and psychomotor abilities (e.g., flight control precision, reaction time, perceptual speed).

Comparisons between experienced and pilots and aviation candidates were used to develop classification formulas for each of the aircraft categories (Intano, Howse, & Lofaro, 1991b). Applying these classification formulas to 686 graduates from Initial Entry Rotary Wing training, Intano, Howse, and Lofaro (1991a) found that use of their battery could improve actual helicopter assignments beyond predictions based on initial flight grades alone.

Selection Instrument for Flight Training

In 2004, a number of concerns stimulated a push for development of a new selection instrument to replace the AFAST. There was a concern that over time the validity of the AFAST had declined. There was also a perceived need for a battery that would relate more closely to changes in Army aircraft and Army aviation doctrine. Thus, ARI began a program to develop the Selection Instrument for Flight Training (SIFT; Katz, 2006). The SIFT was to be computer administered and was to take maximum advantage of instruments already in use to minimize development costs (Paullin, Katz, Bruskiewicz, Houston, & Damos, 2006).

A job analysis and literature review were initiated to identify relevant knowledge, skills, and attributes, and measures of these, associated with successful aviation performance. A variety of cognitive, non-cognitive, and perceptual measures emerged from this approach. These measures were administered to 240 Army aviation students and linked to a variety of training performance measures (Katz, 2006).

A cognitive measure, the Navy's Aviator Selection Test Battery (ASTB), served as the starting point for determining which measures should be included in the SIFT after this preliminary validation. Beyond that, "three criteria were assessed: 1) the amount of incremental validity provided over the ASTB cognitive tests, 2) administration time, and 3) Potential logistical or administrative issues that could make implementation difficult (Katz, 2006, para. 7)."

The ASTB measured "Reading Comprehension, Mathematical Ability, Mechanical Comprehension, Spatial Apperception, and Aviation and Nautical Comprehension" (Katz, 2006). Added to this were "the *Army Aviation Information Test*, and a composite score of *Perceptual Speed and Accuracy* subtests (Katz, 2006, para. 7)." The Army Aviation Knowledge Test was designed as an indirect measure of motivation to become an aviator, based on the assumption that such a motivation would lead to the acquisition of aviation knowledge. There were two timed Perceptual Speed and Accuracy subtests, a Hidden Figures test and a Simple Drawings measure which required examinees to determine which of five drawings was unlike the other four (Katz, 2006).

Summary

There are a number of significant characteristics regarding past and present Army aviation research. One involves the recognition of the value of interservice coordination. Just as early Army research made important contributions to the Air Force's pilot selection program, more recently the Army has constructively exploited the research developments of other services.

The second characteristic is the multi-dimensional nature of the selection and classification tests used. From the beginning of

this research history, there was a recognition that there was no single test that would adequately predict aviation success. Cognitive ability, personality/background, psychomotor and spatial tests have all been proposed to cover the multiple skills and characteristics believed to be associated with the complex and challenging tasks associated with flying aircraft, whether they be fixed wing or rotary wing.

Use of psychomotor tests has historically been constrained by the costs associated with the special equipment needed to fully assess psychomotor ability and the difficulty of maintaining reliability in the measurement of such ability. These factors have held back the inclusion of psychomotor measures in the recently developed SIFT battery.

One area that seems ripe for further development is that on non-cognitive testing. A number of non-cognitive tests were considered for inclusion in the SIFT, although none were incorporated. Measures of personal traits were included in the multi-track aviation classification battery. A number of recent studies have explored personality characteristics associated with aviation success (e.g., Grice & Katz, 2006, 2007). The recent successful use of non-cognitive measures such as the Assessment of Individual Motivation (AIM) and the Tailored Adaptive Personality Assessment System (TAPAS) for enlisted selection, detailed elsewhere in this book, also suggest that interest in personality measurement for aviation selection will be an important factor in future research.

Since World War II, Army aviation classification research has been tied to assignment to one of a variety of helicopter types. The idea that requirements for successful operation of one type of helicopter might be so different than those for another is a fascinating proposition, and the fact that this research has

yielded useful findings is encouraging. Given that helicopter design, operating characteristics and missions are likely to continue to evolve, it would be worthwhile to try to identify some general conclusions about which aspects of each of these are likely to have the greatest impact on classification decisions.

References

Brown, W. R., Dohme, J. A., & Sanders, M. G. (1982). Changes in the U. S. Army aviator selection and training program. *Aviation, Space, and Environmental Medicine, 53,*1173-1176.

Cross, K. D. (1997). *Current state of Army aviator selection* (Research Note No. 97-22). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Davis, F. B. (Ed.) (1947). *The AAF Qualifying Examination. Army Air Forces, Aviation Psychology Program Research Reports* (Report No. 6). Washington, DC: Government Printing Office.

Drucker, A. J., & Kaplan, H. (1966). *Identifying successful pilots through research.* U.S. Army Aviation Digest, 29-32.

DuBois, P. H. (Ed.) (1947). *The classification program. Army Air Forces, Aviation Psychology Program Research Reports* (Report No. 2). Washington, DC: Government Printing Office.

Flanagan, J. C. (Ed.) (1948). *The Aviation Psychology Program in the Army Air Forces. Army Air Forces, Aviation Psychology Program Research Reports* (Report No. 1). Washington, DC: Government Printing Office.

Grice, R. L., & Katz, L. C. (2006). *Personality profiles of experienced U. S. Army aviators across mission platforms* (Tech. Rep. No. 1185). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Grice, R. L., & Katz, L. C. (2007). *Personality profiles of U. S. Army Initial Entry Rotary Wing students versus career aviators* (Tech. Rep. No. 1208). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Henmon, V. A. C. (1919, June). Air service tests of aptitude for flying. *Journal of Applied Psychology, 3,* 103-109.

Intano, G. P., Howse, W. R., & Lofaro, R. J. (1991a). *Initial validation of the Army aviator classification process* (Research Note No. 91-38). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Intano, G. P., Howse, W. R., & Lofaro, R. J. (1991b). *The selection of an experimental test battery for aviator cognitive, psychomotor abilities and personal traits* (Research Note No. 91-21). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Intano, G. P., & Howse, W. R. (1992). *Predicting performance in Army aviation primary flight training* (Research Note No. 92-06). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Katz, L. C. (2006). *Find the "Right Stuff": Development of an Army aviator selection instrument.* Presented at the 25th Army Science Conference, Orlando, FL.

Melton, A.W. (Ed.) (1947). *Apparatus tests. Army Air Forces, Aviation Psychology Program Research Reports* (Report No. 4). Washington, DC: Government Printing Office.

Paullin, C., Katz, L., Bruskiewicz, K. T., Houston, J., & Damos, D. (2006). *Review of aviator selection* (Tech. Rep. No. 1183). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Thorndike, E.L. (1919). Scientific personnel work in the Army. *Science, 49*, 53-61.

Thorndike, R. L. (Ed.) (1947). *Research problems and techniques. Army Air Forces, Aviation Psychology Program, Research Reports* (Report No. 3). Washington, DC: Government Printing Office.

Yerkes, R. M. (1919, March). Report of the Psychology Committee of the National Research Council. *Psychological Review, 26*, 83-149.

PART IV

PERSPECTIVES ON THE PAST AND LOOKING TO THE FUTURE

CHAPTER 14

PERSPECTIVES ON THE PAST AND LOOKING TO THE FUTURE

Michael G. Rumsey
U.S. Army Research Institute for the Behavioral and Social Sciences

The Army's selection and classification advancements described in this book are truly remarkable. They encompass some of the major historical advances in the history of all selection and classification research, including the development of group testing techniques in World War I and the ground-breaking validation research in Project A and the Career Force. Contrary to what some may think, however, the major challenges in selection and classification research are not behind us. We are entering an era where the demands on such research are escalating, and major advances are still needed.

Recent Army future-oriented work, particularly a project known as Select21 (Knapp & Tremble, 2008), has identified future demands that place heavy emphasis on both cognitive and noncognitive attributes. Future soldiers will need to deal effectively with sophisticated technology, particularly information technology, will need to be highly adaptive, and will need to deal with various sources of stress. These trends would seem to suggest the need for higher standards. Yet, higher standards would reduce the supply of qualified applicants in a difficult recruiting environment. This is a serious

dilemma, one that calls for focused attention on ensuring that the best possible selection and classification system is in place, and that this system is integrated with the Army's personnel, training, and mission needs.

Defining and Measuring Success

Types of Success

At the most basic level, the purpose of selection is to identify those who will succeed on the job. Success is defined in terms of both performance and retention. The Army's primary tool for predicting successful enlisted performance has been the cognitively-based Armed Services Vocational Aptitude Battery (ASVAB) and its primary tool for predicting enlisted retention has been high school graduation. Often these two selection tools are referred to as indicators of "quality."

Beginning with Project A, researchers began to redefine the meaning of "quality." In part, this comes from a better understanding of what constitutes performance. Project A demonstrated that performance is more complex than the completion of training or the accomplishment of specified tasks, which we have summarized under the broad term "can do" performance. It demonstrated that there is a separate major component that we have labeled "will do" performance, and that in order to predict this component, we need more than a cognitive test battery.

The five performance dimensions identified in Project A can serve as a starting point for a better understanding of the complexity of soldier performance, but cannot yet be accepted as the final word on the subject. One cannot develop a performance measure for a job unless one understands what are the critical requirements of that job. The process of identifying

such requirements is typically termed job analysis. The Project A performance measures were built from an analysis of Army jobs that was unusually comprehensive, but in most respects was not intended to be particularly innovative. The underlying philosophy was to take advantage of job analysis information that the Army had already collected to avoid duplication of effort. The approaches used were generally based on a model of a fairly well structured and stable job, a model which becomes increasingly questionable in an increasingly fluid and dynamic world. They were not designed to explore the complex interrelationships between behaviors occurring at earlier times, at the same time, or at later times, or the interactions between individual, group, and organizational performance. The model of performance developed in Project A represented a great leap forward in terms of our basic understanding of performance, but there remains more to be done.

Understanding the Job: Job Analysis

Developments in understanding performance need to go hand in hand with developments in job analysis. The identification of motivational elements of performance in Project A parallels research findings elsewhere that suggest that jobs consist of both task components and components that are not formal requirements of the job, but contribute to organizational effectiveness (Borman & Motowidlo, 1993). These evolving concepts suggest that there is much about work that is not fully understood and requires closer examination. A critical need for future research is to advance the science of job analysis. One objective of such research is to improve the efficiency of the analysis process. So long as the examination of each job is viewed as more or less independent of every other job analysis, efficiency will be difficult to achieve. One promising alternative

approach focuses on building a taxonomy of job components that has applicability across jobs (for an example of such an approach, see Peterson, Mumford, Borman, Jeanneret, & Fleishman, 1999). Then, the problem becomes less one of discovering the requirements of each job from scratch, and more one of determining which components of an existing taxonomy apply to a particular job.

Changes in missions, equipment, doctrine, and technology ensure that jobs will not remain entirely stable. Jobs are frequently expanded, eliminated, added, or combined with other jobs. This suggests that we need to have ways of looking at jobs that are sufficiently flexible to allow adaptation to structural changes. This need helps account for the recent growing popularity of a concept known as competency modeling (Shippman, et al., 2000). Competencies are understood to encompass broad units of work, which are grouped by individual capabilities rather than specific organizationally determined job components. While such a flexible approach may indeed facilitate job restructuring in the future, greater clarity in how competencies are to be identified, defined, and combined will be necessary before effective use of such an approach will be possible.

A longer-term objective is to develop an approach that fully recognizes the complexity of jobs. Job components such as tasks are often not in reality discrete and independent, but rather fluid and interactive. The behaviors that comprise these components interact with one another, with organizational objectives, and with behaviors that others in the organization exhibit. A broader concept than job analysis is needed to fully recognize the complexity and gestalt represented by a job. Work analysis (Sanchez, 1994) and job synthesis (Rumsey, 2007)

have been proposed as possible alternative concepts. A more realistic portrayal of job requirements based on a greater appreciation of the complexity of the job will lead to more sophisticated and realistic measures of performance.

Measures of Success

While a better understanding of the job, whether through job analysis or synthesis, is an essential precondition to the development of appropriate performance measures, it is not by itself sufficient to achieve that goal. How does one represent and measure the full complexity of the job?

Improved performance measurement will require both advancements in existing methods and exploration of new methods. Most performance measures fall into one of two broad categories—assessment of actual job performance or of simulated job performance. Examples of methods measuring actual job performance are ratings and administrative measures. Considerable research has been devoted to improving ratings, and the methodology can be described as relatively mature. Whereas ratings are often designed for research purposes, administrative measures by definition represent information collected by the organization. The challenge regarding use of administrative measures for research is to relate this information to important dimensions of performance. Project A provided an excellent model of this approach (see Knapp, Campbell, Borman, Pulakos, & Hanson, 2001).

Perhaps the most fertile ground for new research on performance measurement methods involves the development of new simulation approaches. The advantage of such approaches is the control they afford to the researcher, who can decide which aspect or aspects of the job he or she wishes to represent, and can control the pace and manner in which these

aspects are presented. Important job requirements that might only be presented infrequently under natural conditions, and thus might not normally be observed by others, can be selected as the focus of a simulation. While simulations in the training world of the Army are often associated with complex, expensive arrangements of individuals and equipments, as in war games, simulations as we discuss them here may be of a much more limited scope.

The hands-on tests used in Project A might be viewed as a mix of observation of actual performance and a simulation of such performance. They measured actual tasks performed on the job, were conducted on the job site, and used the equipment used on the job. However, since they were performed for research purposes, not to meet operational job requirements, they did not completely represent job performance. Since they focused on discrete tasks, not the full complexity of job behavior, they might be considered a very rudimentary form of simulation.

Another type of simulation is the situational test, which we have seen in the history presented in this book used sporadically. That is to say, the individual is presented with a situation or series of situations and evaluated on his or her reactions. Often these situations are combined into complex assessment centers. Situational tests attempt to replicate aspects of the job the individual will be expected to perform. Because of the expense involved, situational tests have seen limited use. The challenge is to match the realism of the situation with the sophistication of the measurement applied. It may be difficult to isolate and measure the particular behaviors associated with successful performance in a situational exercise. This is a problem to be addressed in future research on this method.

A paper and pencil form of a situational test, known as a situational judgment test, has become increasingly popular in recent years. Situational judgment tests represent both a lower fidelity (Motowidlo, Dunnette, & Carter, 1990) and lower cost approach to evaluating an individual's judgment of the best response to a situation. Because they contain verbal material, they clearly have a cognitive component, but it is less clear what other individual qualities are being measured. Sternberg (1994) has suggested that situational tests he has developed, which he calls tacit knowledge tests, measure knowledge that is acquired through experience. He further equates performance on these tests with a special kind of intelligence, which he has labeled practical intelligence. While this claim is controversial, the potential of such tests will likely receive further examination in the future.

A third means of presenting situations is by interactive video (Salgado, 1999, discusses use of video technology in selection). Although there is some history in the Army of the use of video for assessment (e.g., Schroeder, Dyer, Czerny, Youngling, & Gillotti, 1986), the sophistication involved in current methods makes early uses now seem relatively primitive by comparison. Early examples were not interactive—the situation was presented, then the examinee indicated which response from a set of alternatives was most appropriate. Then a new situation was presented. Now, the examinee can be a participant in an ever-evolving story—giving responses to stimuli which have an impact on how the story develops. The situations can be presented using advanced technology that gives them a high degree of realism. Here, the situational tests truly merge into simulations, with no apparent distinction between the two.

Computer simulations may represent a level of abstraction away from a real world context, but can compensate for this by the degree they can represent the complexity of that world. They can generate "microworlds" (see Kluge, 2008, for a discussion of this technique) which encompass many of the environmental variables and much of the dynamism found in the world of work. This can add to the credibility and attractiveness of the test to the examinee. They can allow for the emergence of adaptive individual behaviors that could not easily be observed in other contexts. However, the sophistication of measurement of individual performance in simulations has not yet caught up with the sophistication of the microworlds themselves (Kluge, 2008). Thus, further research is needed before the potential of simulations for performance measurement purposes can be realized.

Future research will need to recognize and further define the complexity of performance. Single measures of performance, whether they be proficiency-oriented, such as job knowledge or hands-on tests, or motivationally oriented, such as ratings or records of accomplishment, cannot by themselves provide a basis for determining the appropriate set of selection and classification tools. Soldier performance is a complex mix of proficiency and motivational dimensions. It cannot fully be represented by a hands-on measure that shows the individual's capability to perform a single task at a single moment of time when that individual is aware that he or she is evaluated. Nor can it simply be represented by a rating from an observer that provides a broader perspective of the individual's performance over time, but must depend on impressions and memory rather than a focused evaluation of particular slices of behavior. Given the challenges and costs associated with a full evaluation of job performance, shortcuts will likely need to be considered for specific purposes,

but such shortcuts should be taken with a full recognition of the trade-offs involved.

Predicting Success

Types of Attributes Measured

Noncognitive Attributes

Personality. At some level, there has always been recognition that performance is the product of more than cognitive ability. The Project A findings focused more attention on those remaining characteristics of the "whole person" that drive the motivational component of performance. They highlighted two major questions: what are these characteristics, and how might they be measured?

Much of the Army research on noncognitive measures since Project A has focused on personality characteristics. Personality dimensions are typically described in terms of the "Big 5": Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience (Goldberg, 1994). The seven Assessment of Background and Life Experiences (ABLE) and the six Assessment of Individual Motivation (AIM) dimensions incorporate modest deviations from this list.

The identification of the Big 5 dimensions has been credited with advancing personality theory from a collection of poorly organized personality characteristics to a more disciplined structure (e.g., Helton & Street, 1992). However, there is much more to be understood about this structure. While undoubtedly a useful means of describing broad dimensions of personality, it is also imprecise. Likely, it will need to be refined and elaborated over time. The task of aligning these broad dimensions with more specific components of personality, often referred to as facets, is only partially complete.

Many have argued that closer attention to specific personality traits will give more precise measurement than the more global dimensions of the Big 5 (e.g., Schneider, Hough, & Dunnette, 1996). Recently, Army research has begun to explore facet-oriented measures. Such a measure, the Tailored Adaptive Personality Assessment System (TAPAS), has been shown to predict both performance and attrition of Army soldiers (Heffner, White, & Owens, 2010), and is now being administered to applicants for Army enlistment. The challenge of finding the right balance between breadth and specificity has been referred to as the bandwidth-fidelity trade-off. Broader coverage of the personality bandwidth is believed to extend the range of behaviors that can be predicted, but with lower precision. More focused coverage of a portion of this bandwidth leads to more limited, but more precise prediction (Vickers, Hervig, & Booth, 1996). Thus, consideration of what one is trying to predict should affect the breadth of the personality measure used.

Future research on personality measures will also have to address the issue of positive self-representation; otherwise known as faking. The question of whether faking is a major problem in personality assessment has been a controversial one, with some offering evidence that faking concerns have been exaggerated (e.g., Schwab, 1971). However, the question is not easily settled. The degree to which it is a problem will vary according to such circumstances as whether the testing environment is a high stakes one, the number of people who are tested, and the extent to which coaching is a potential issue. Those participating in the enlisted accessions process have a powerful incentive to fake, and the results indicate that faking in this environment is a problem. It was soon discovered that the positive results with ABLE in the research context did not

necessarily translate to positive outcomes in an operational environment. The most recent version of AIM has been found to be faking-resistant even in an operational environment (White, Young, Hunter, & Rumsey, 2008), but it would be premature to declare victory in the war on faking. Additional research will be needed to further refine defenses against the faking threat.

It was stated above that success is defined in terms of two types of outcomes—successful performance and retention. The Army has historically used separate tools for each—ASVAB to predict performance, and evidence of a high school diploma to predict retention. We have found that personality measures can be used to predict both performance, specifically "will do" performance, and retention. This enhances their value as selection tools. AIM has been successfully used to predict retention among non high school diploma graduate enlisted soldiers (White, et al, 2008). Now, there is growing interest in the use of personality measures in the selection process to predict officer retention. As noted in Chapter 11, research on officer selection has waxed and waned. Although personality measures have been used in officer selection before, they are essentially absent from the selection process now. In times of constrained personnel resources, the need to consider retention in the officer selection process becomes critically important. We have entered such times, and so long as these constraints persist, we can expect that the need for officer selection research based on projected retention will receive considerable attention.

Biodata. Personality measures such as the AIM offer one approach to measuring characteristics relevant to future performance. Biodata measures, which ask questions about the individual's background, constitute another approach. While the questions may differ, biodata measures may tap into some of the

same characteristics as personality measures. Because the combination of personality and biodata measures may yield more information about the individual than either approach alone, there is value in continuing to pursue both approaches. There is a basis for believing that biodata questions can be particularly resistant to faking, as one might be presumed to be less likely to try to mislead about what one has done than what one believes, particularly if one perceives that one's behavior has been observed (Mael, 1991). One potentially fruitful avenue for research might involve developing strategies to reinforce this potential advantage for biodata measures.

Cognitive Attributes

When one has a cognitive measure that predicts job proficiency as well as the ASVAB, one must consider how much improvement is really possible. To answer this question, we need to return to the issues of job analysis and performance measurement. The job analyses conducted in Project A, while extraordinarily comprehensive, generated performance measures that were relatively particularistic. The spatial and psychomotor measures examined in that project were able to add little to the prediction of such outcomes. However, they were found to add considerably to the prediction of gunnery tests that were not part of the Project A criteria (Walker & Rumsey, 1994).

The Army's Select21 project suggested that we are likely to face a future characterized by a relatively high degree of ambiguity and change. To the extent future Army researchers are able to accurately represent these characteristics in their measures of performance, they might find that cognitive measures that have added little to the ASVAB in the past emerge as more powerful predictors. These might include

measures of speed of information processing, short-term memory, cognitive flexibility, and cognitive complexity.

Changing future requirements might also stimulate interest in new types of information tests, beyond those currently represented in the ASVAB. Computer literacy is already receiving considerable interest in the Joint Service context (Drasgow, Embretson, Kyollonen, & Schmitt, 2006), generating questions about how this capability might be defined and measured. There is some question about what computer literacy is—an ability, possession of knowledge or understanding about computers, or something else. Future research will need to explore this question further and determine whether inclusion of a test on computer literacy can improve the current classification efficacy of the ASVAB.

As our nation becomes increasingly culturally diverse, the question of whether the ASVAB adequately represents this cultural diversity becomes an increasingly pressing issue. There is some concern that the greater the use of the English language in the test, the greater will be the handicap for those who are not proficient in this language. Thus, there is growing interest in the use of non-verbal tests for selection and classification. One such test, Assembling Objects, was found to be particularly successful for predicting performance in Project A. Although a spatial test, it was found to have less adverse impact on females than other spatial tests (Russell & Peterson, 2001). Further research may well consider how Assembling Objects and other non-verbal tests might be used in enlisted selection and classification.

Measurement Methods

Computer Testing

Already much of the enlisted selection testing is done by computer. The increased use of computer testing is a trend that is likely to increase, because of such advantages as time savings and flexibility. Let us briefly discuss each of these in turn.

Computer administration can reduce testing time by making possible the use of a procedure called "adaptive testing." Adaptive testing means that the test items are not the same for everyone—they are adapted to each individual. The ASVAB is already administered to the bulk of applicants in a computer adaptive format. On an ASVAB test, if an individual does well, the items get progressively more difficult; if he or she does poorly, the opposite happens. This method helps pinpoint the individual's location on a continuum with fewer items than a test that is not tailored to the individual. However, an area ripe for additional research involves adaptive testing in the personality domain, which the Army has now taken to the level of an operational test in its work on TAPAS (Knapp, Heffner, & White, 2011). This approach allows broader coverage of personality dimensions in a limited testing period. Again, the individual is measured on a continuum. For example, for the personality dimension "work motivation," we would try to find where the individual fits on a continuum from very low to very high.

Computer testing provides greater flexibility by making possible the use of approaches not possible in a paper and pencil context. Administration of certain psychomotor and spatial tests, such as the Target Shoot test in Project A, are only possible through use of a computer. Computer testing also allows for more precise control of testing time than is possible

with a paper and pencil test, whether it be time for the total test, for an item, or for a single response within an item.

Simulations

Typically, selection measures are focused on particular personal characteristics that are believed to be important for successful job performance. However, in some cases more elaborate assessment procedures are used to more directly determine how an individual can be expected to perform in actual job situations. Earlier in this chapter, we discussed various types of simulations used to measure job performance. Some of the same types, including situational tests, situational judgment tests, computer simulations, and interactive videos, have also been used for selection (Salgado, 1999). For this purpose, such approaches cannot rely on job content that candidates would not be expected to know. However, if they incorporate relatively general types of situations, such as presenting a talk or planning a meeting, they can tap into combinations of attributes that may be more indicative of an individual's likely future performance than single-attribute measures. Because of the expense associated with such measures, they have not been extensively used in the past. Future research might be devoted to making them both more cost-efficient and more closely linked to relevant job characteristics than currently available versions.

Interview

Another method that will likely require future consideration is the interview. While often discredited as a poor means of assessing individual talent, the interview has gained credibility as the sophistication of the technique has matured. Greater structure has improved standardization and ensured more focus on relevant content. Here again, the situational approach has been explored (Salgado, 1999). Presenting the individual with a

situation and asking how he or she would respond represents an oral alternative to a written situational judgment test.

Credibility of the testing tool has increasingly been recognized as an important component of the selection process (Salgado, 1999). One of the reasons behind the computerization of the ASVAB was that computer testing is perceived to be consistent with the concept of a technologically sophisticated military. Future research on computerized testing, simulations, and other approaches can be expected to show continual concern with the credibility issue. To the extent that tests are designed in such a way as to be seen as relevant to the selection and classification mission, their credibility will be enhanced.

Challenges for Classification Research

Equally important to the issue of selection is that of successful matching of persons to jobs, or classification. Successful classification requires the capability to discriminate requirements from one job to another, the capability to develop individual difference measures which are closely linked to differences between jobs, and the capability to develop procedures which capture the full potential of these individual difference measures to achieve the greatest possible benefit to the organization.

Job Analysis for Classification

Discriminating requirements from one job to another leads us back to the job analysis and performance measurement issues discussed earlier in this chapter. The ability to determine the comparability of jobs is critical to the ability to conduct classification research. Yet, the cost of developing separate outcome measures for over 180 jobs is beyond the budget of most research projects. One possible alternative is to group jobs

together on the basis of common characteristics, and develop outcome measures that have applicability to multiple jobs. If we start with a complete taxonomy of such characteristics, we will have a starting point for such a comparison. Work has already begun on examining the applicability of an existing taxonomy, known as O*NET (Peterson, Mumford, Borman, Jeanneret, & Fleishman, 1999), to Army jobs. While the preliminary findings are encouraging, there is much to be done before this kind of approach can be considered the solution to the problem we are facing in developing adequate job analysis and performance measurement tools for classification purposes.

The question of whether a job characteristic that appears in multiple jobs really is the same in one job as in another is a major issue. Let us say the characteristic is that the job involves leading people. Does that characteristic have the same meaning in an Infantry job as it does in a Motor Vehicle Repair job? Will it relate to the same individual difference characteristics in both jobs? Another issue is the level of detail needed to adequately represent that characteristic. It may need to be described in general terms to be applicable to multiple jobs, but it may need to be presented in specific terms to assess how the requirements for one job differ from those in another. Thus, while the taxonomic approach is well worth pursuing, it may or may not provide a satisfactory solution to the problem of developing cost-efficient performance measures for classification purposes.

Individual Difference Measures for Classification

Next, we come to the second requirement for successful classification—the capability to develop individual difference measures which are closely linked to differences between jobs. The ASVAB is the primary tool used for this purpose at present. It has been shown to be a strong predictor of job proficiency,

but has not been found to be as useful for predicting "will-do" performance. Personality measures have been shown to relate to such performance. Future research is needed to explore how noncognitive measures can be used to better determine an individual's suitability for a particular job. As noted above, computerized adaptive testing opens the door to cover a relatively large number of personality characteristics. The more discrete components of an individual's personality we can measure, the more likely it is that we will be able to identify those that have unique relevance to particular jobs. What personality characteristics are associated with success as a truck driver? As a member of the military police? We need to find the best combination of aptitudes and personality characteristics for a particular job, to avoid misclassifications on either basis.

Interest measurement may well also play a major role in future classification research. In Select21, we found ways of measuring interests that both provided a basis for matching interests to jobs and also demonstrated linkage to retention intentions. Because it is difficult for candidates to know enough about a particular job to know whether or not they might be interested in it, measurement of interests in activities or attributes associated with such a job is a useful indirect way of measuring interest in that job. Given that retention in the Army is becoming an increasingly important outcome, any change to job placement that can enhance retention is likely to be highly valued.

Developing a Classification System

Now, we consider the third requirement of successful classification—the development of systematic procedures that link the attribute measures to the jobs. With classification one is not just deciding which individuals are qualified for which jobs. Each classification decision involves comparing the benefit to

the Army of one job placement relative to the value of every other job match possibility for that individual. A person may be well qualified for one job, but, because of the other individuals also qualified for that job, may serve the Army better in another job. With over 180 military occupational specialties in the Army and approximately 80,000 accessions a year, the mathematical challenges associated with evaluating these different alternatives are enormous. Research over the past 20 years has provided two tools for meeting these challenges. The first is Zeidner and Johnson's (1994) Differential Assignment Theory, which determines which weights should be applied to each test for determining the quality of the match of an individual with a job. The second is the Enlisted Personnel Allocation System (EPAS), which uses these weights and other factors in determining the priority of job opportunities that should be offered to each candidate. The joining of DAT with EPAS represents a great step forward by providing a solution that recognizes the full complexity of the classification problem and sophisticated statistical techniques to address it.

The future of EPAS, and by extension full utilization of DAT, will likely depend upon the extent to which EPAS can be reconciled with operational constraints. Certain jobs are designated priority jobs, to be filled first. If the urgency of filling these jobs is judged to be more important than maximizing the fit between the individual and the job, the potential of EPAS will not be realized. EPAS' benefits are likely to be underestimated if users lack confidence that, if a critical job is not filled immediately, it will be filled soon enough by this system. Future research in this area is likely to pursue issues associated with user acceptance, system constraints, and individual preferences.

Special Assignment and Mission Issues

Earlier, we equated classification with matching of persons to jobs. Thus far, we have examined the broad problem of matching entry-level personnel to over 180 Army jobs. An individual may also be matched with a particular type of assignment or mission that may or may not be linked with a particular job. These types of decisions entail certain research challenges that are somewhat different from those discussed earlier and merit our attention now.

Within this category are assignments or missions that involve teamwork. For such assignments, one needs to take a broad view of the job and consider which attributes contribute to effective team functioning as well as to effective task performance. While there is an evolving field addressing characteristics that contribute to effective teamwork (Brannick & Levine, 2002), this field is not sufficiently blended with selection research.

Also within this category are assignments that require special skills, such as Special Forces. While there is a solid history of research on Special Forces selection, the research reported in this book suggests that the need for special skills is likely to increase in the future. These skills may both include and go beyond those currently required. Continued probing of future requirements is needed to insure that our selection procedures keep pace with these requirements.

Two types of missions deserve special attention—combat and peacekeeping. One recurring concern involves the link between peacetime and combat performance. Despite the criticality of combat performance, our selection measures are typically validated on peacetime performance measures. There are a number of practical obstacles to conducting combat

performance assessments, but progress in this area could pay major dividends. The challenge of predicting combat success is complicated by the wide diversity of combat missions and concomitant diversity in determinants of successful individual combat performance. Are the behaviors associated with success in World War II the same as those associated with success in the Vietnam War, or in more recent combat operations in Afghanistan and Iraq? Are the behaviors that led to success in the early phases of the Afghanistan and Iraq conflicts the same as those associated with success in 2010? Of course, the key question is what types of behaviors will be important in future, not necessarily those which were important in past, operations.

Peacekeeping has become an increasingly prominent role for the Army. Yet, there has been limited research on the attributes that lead to effective performance in this role (for one such rare venture, see Mael, Kilcullen, & White, 1996). As with combat, there are a variety of peacekeeping environments, from less to more hostile and from less to more involvement in host nation government operations (e.g., civil affairs), suggesting that the identification of critical attributes for effective peacekeeping performance will require thoughtful analysis.

Selecting for Future Advancement

At the outset of this chapter, it was noted that selection is designed to support two major outcomes: successful performance and retention. There are a variety of perspectives that can be taken with respect to the type of performance that is of interest. One perspective is that prediction of training performance is sufficient, since completion of Initial Entry Training is the first job requirement that must be satisfied. The utility of this approach will depend to a considerable degree on how well training success predicts on-the-job success. Greater

attention is needed to developing measures of training success that contain as many elements of later job success as possible.

A more ambitious perspective suggests that one need also be concerned with the soldier's performance in his or her first unit of assignment. While this perspective leads to research that is more costly and generally more time consuming than one that focuses only on training, it is more directly associated with fulfilling the Army's ultimate need. Beginning with Project A, the Army has conducted a number of projects that incorporated measurement of on the job performance.

However, there is an even more ambitious perspective that can be advanced. When one selects which civilians meet the standards for entry into the Army, one is also creating the pool from which will be drawn those who advance to higher ranks. With limited exceptions, promotion in the Army is from within. Not everyone who enters is expected to advance, but there must be a large enough pool of high potential individuals within the initial cohort to ensure that the higher ranks will ultimately be filled with qualified personnel. Attributes associated with success at initial entry are not entirely identical with attributes needed for success at higher levels. Ability to deal with complexity, for example, has been associated with success at executive levels (Jacobs & Jaques, 1991). Thus, development of future selection tools must be guided by the need to balance these two not entirely compatible objectives.

Special Issues

Here, we address briefly a number of other topics in selection and classification that may merit some attention in the coming years.

Validation Strategies

A validation strategy is, in the context of this chapter, a means of assessing the value of measures developed for selection and classification. The most direct way of conducting this assessment is to determine the relationship between these measures and some outcome, which becomes known as the criterion. However, this approach can be expensive and may take a number of years. A recent review of the ASVAB (Drasgow et al., 2006, p. 46) recommended that the services "relax the requirement for criterion validity of new measures." The reviewers noted that there is other evidence that can be considered in determining whether a measure serves its intended purpose, such as an examination of how appropriate the content of the measure is with respect to what it is trying to measure. We can expect that future research will consider alternative strategies for assessing the value of new measures, but we can also expect that criterion validity will continue to play an important role in Army selection and classification research.

Internet Testing

The selection and classification system exists to help the Army meet its manpower needs. Essential to this goal is making the process as user-friendly as possible. Currently, the screening process requires the individual to travel to some testing station to qualify for entry. Naturally, the Army would like to minimize the burden on the examinee in order to encourage the maximum number of people to apply. This suggests that the Internet may play a role in future selection and classification. There are a number of practical considerations that must be addressed before this role can be better defined. The greatest barrier to Internet testing for record purposes is security. How do we

ensure that the person taking the test is the actual candidate and not someone else? How do we ensure that the Internet is not used to obtain the test questions and ultimately compromise the test? How do we keep Internet testing from becoming an "open book" exam rather than a test of the candidate's knowledge and potential? Until these questions can be satisfactorily answered, there will need to be limits on how the Internet can be used for selection and classification testing.

Integration of Selection with Initial Training and Assignment

As noted above, one purpose for selection is to identify those who will succeed in initial training. However, the selection issues pertaining to training go beyond the prediction of success or failure. To what extent should information collected at entry be used to facilitate later training? Should entry tests be used to identify training needs? How well can they serve this end without compromising their primary purpose? To what extent can scores on initial selection tests be used in assignment to training regimens and to particular job assignments? What kind of automated database system needs to be developed to ensure optimal use of selection and classification information? Attention to questions such as these can enhance the overall organizational value of the selection system.

Conclusions

After the completion of Project A and Career Force, some believed that there was little need for future selection and classification research in the Army. There was much that was learned in Project A, both about the nature of performance and about the manner in which measures of various types of attributes related to different components of job performance.

But far from sounding the death knell of selection research, Project A actually stimulated a number of new research initiatives. This impact was seen most profoundly in the realm of noncognitive measures. Project A offered hope that such measures could be used to provide a more comprehensive assessment of individual capabilities than could cognitive measures alone. But noncognitive measurement had not advanced to the point where it could easily be inserted into the enlisted selection process. Since then, much progress has been made and the Army has crossed the threshold of using noncognitive measures in selection. However, in an ironic way, this success has actually increased the need for further developments in this area, so that the full potential of such measures can be realized.

The need for further research on cognitive measures is less apparent, but no less real. While the usefulness of such measures has been demonstrated many times, the need to ensure that they are maximally linked to relevant job requirements remains. Does their heavy reliance on verbal content limit their relevance for identifying successful applicants for jobs that do not require mastery of verbal skills? Are there alternative methods to those currently employed in the ASVAB that better match individual capabilities with job requirements? Are there cognitive measures that can better address future job demands than the ASVAB? These are questions that merit further research.

Improved selection and classification is inevitably linked to improved understanding of performance. Thanks to Project A, the complexity of soldier performance has been revealed with a clarity never before seen. However, the challenges of developing both realistic and cost-effective measures of

performance, and of distinguishing performance requirements from one job to another, remain.

The first step towards improved performance measures lies in improving upon current methods of understanding individual jobs. There is a multiplicity of job analytic methods available, but none which fully provides a comprehensive understanding of how jobs either resemble or vary from one another, and which can do so in a sufficiently efficient manner.

The second step towards improved performance measures addresses the measures themselves. As our understanding of the dynamics of successful job performance is enhanced, the inability of current approaches to fully represent these dynamics will become more manifest. One promising path to better representation of the complexity of job performance is by means of simulations. There is much room for improving the measures used in simulations, and that is one direction future research may take us. Another is to build upon the progress already made toward building low cost simulations that target limited aspects of job performance.

Much is demanded of the tools used for selection and classification. They are expected to identify those who will succeed and those who will fail in a variety of multi-faceted jobs, and they are expected to do so with a high degree of precision through the use of tests that are often administered at a single time and place. The tests must usually be relatively short and cheap. Because of these high expectations, research on these tests is not a short-term exercise. There is not likely to be a single test that can adequately serve all needs. Multiple methods will be needed, and considerable attention will need to be given to how best to combine these methods for various purposes.

The Army has come far toward meeting its objectives. It is on the verge of making whole person assessment a reality, and thereby transforming the future direction of selection and classification research. The power of the new tools to be provided to the Army will generate a new appreciation of their ability to bring out the best among those who constitute the most precious resource an organization can offer—its human resource.

References

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance*, *10*, 99-109.

Brannick, M. T., & Levine, E. L. (2002). *Job analysis*. Thousand Oaks, CA: Sage.

Drasgow, F., Embretson, S. E., Kyllonen, P. C., & Schmitt, N. (2006). *Technical review of the Armed Services Vocational Aptitude Battery* (Final Rep. No. 06-25). Alexandria, VA: Human Resources Research Organization.

Goldberg, L. R. (1994). Basic research on personality structure: Implications of the emerging consensus for applications to selection and classification. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 247-260). Hillsdale, N.J.: Erlbaum.

Heffner, T. S., White, L., & Owens, K. S. (2010). *Tier One Performance Screen.* Paper presented at the meeting of the Army Science Conference, Orlando, FL.

Helton, K. T., & Street, D. R. Jr. (1992). *The five-factor personality model and naval aviation candidates.* Pensacola, FL: Naval Aerospace Medical Research Laboratory (NTIS No. AD-A260 227).

Houston, J. S., Borman, W. C., Farmer, W. F., Bearden, R. M. (2006). *Development of the Navy Computer Adaptive Personality Scales (NCAPS)* (Tech. Rep. No. 06-2). Millington, TN: Navy Personnel Research, Studies, and Technology Division, Bureau of Naval Personnel.

Jacobs, T. O., & Jaques, E. (1991). Executive leadership. In R. Gal & D. Mangelsdorff (Eds.), *Handbook of military psychology*. Chichester, UK: Wiley.

Kluge, A. (2008). Performance assessments with microworlds and their difficulty. *Applied Psychological Measurement, 32,* 156-180.

Knapp, D. J., Campbell, C. H., Borman, W. C., Pulakos, E. D., & Hanson, M. A. (2001). Population assessment for a population of jobs. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 181-235). Mahwah, NJ: Erlbaum.

Knapp, D. J., Heffner, T. S., & White, L. (Eds.) (2011). Tier One Performance Screen Initial Operational Test and Evaluation: Early results (Tech. Rep. No. 1283). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J., & Tremble, T. R. (2008). Future-oriented experimental Army enlisted personnel selection and classification project (Select 21) (Tech. Rep. No. 1224). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Mael, F. A. (1991). A conceptual rationale for the domain and attributes of biodata items. *Personnel Psychology, 44,* 763-792.

Mael, F. A., Kilcullen, R. N., & White, L. A. (1996). Soldier attributes for peacekeeping and peacemaking. In R. H. Phelps & B. J. Farr (Eds.), *Reserve Component soldiers as peacekeepers* (pp. 29-57). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: the low-fidelity simulation. *Journal of Applied Psychology, 75,* 640-647.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., & Fleishman, E. A. (1999). *An occupational information system for the 21st century: The development of O*NET.* Washington, DC: American Psychological Association.

Rumsey, M. G. (2007, August). *What's in a job?* Paper presented at the meeting of the American Psychological Association, San Francisco, CA.

Russell, T. L., & Peterson, N. G. (2001). The Experimental Battery: Basic attribute scores for predicting performance in a population of jobs. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 269-306). Mahwah, NJ: Erlbaum.

Salgado, J. F. (1999). Personnel selection methods. In C. L. Cooper & L. T. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 1-54). New York, NY: John Wiley & Sons.

Sanchez, J. I. (1994). From documentation to innovation: Reshaping job analysis to meet emerging business needs. *Human Resource Management Review*, *4*, 51-74.

Schneider, R. J., Hough, L. M., & Dunnette, M. D. (1996). Broadsided by broad traits: how to sink science in five dimensions or less. *Journal of Organizational Psychology*, *17*, 639-655.

Schroeder, J. E., Dyer, F. N., Czerny, P., Youngling, E. W. & Gillotti, D. P. (1986). *Videodisc Interpersonal Skills Training and Assessment (VISTA): overview and findings, Volume 1*. Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences (AD A168 287).

Schwab, D. P. (1971). Issues in response distortion studies of personality inventories: a critique and replicated study. *Personnel Psychology*, *24*, 637-647.

Shippman, J. S., Ash, R. A., Battista, M., Carr, L., Hesketh, B., Kehoe, J., Pearlman, K., Prien, E. P., & Sanchez, J. I. (2000). The practice of competency modeling. *Personnel Psychology*, *53*, 703-740.

Sternberg, R. J. (1994). The PRSVL model of person-context interaction in the study of human potential. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 317-332). Hillsdale, NJ: Erlbaum.

Vickers, R. R., Hervig, L. K., & Booth, R. F. (1996). *Personality and success among military enlisted personnel: An historical prospective study of U. S. Navy corpsmen* (Rep. No. 96-15). San Diego, CA: Naval Health Research Center.

Walker, C. B., & Rumsey, M. G. (1994). Application of findings: ASVAB, new aptitude tests, and personnel classification. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 457-466). Hillsdale, NJ: Erlbaum.

White, L. A., Young, M. C., Hunter, A. E., & Rumsey, M. G. (2008). Lessons learned in Transitioning personality measures from research to operational settings. *Industrial and Organizational Psychology*, *1*, 291-295.

Zeidner, J., & Johnson, C. D. (1994). Is personnel classification a concept whose time has passed? In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 377-410).. Hillsdale, NJ: Erlbaum.

TOPIC INDEX

NAME INDEX

Peter F. Ramsberger

Dr. Peter F. Ramsberger is the Manager of the Center for Personnel Policy Analysis within the Human Resources Research Organization (HumRRO) in Alexandria, Virginia. Dr. Ramsberger holds a Ph.D. in Social Psychology from the George Washington University in Washington, D.C.

Nikki R. Wooten

Dr. Nikki R. Wooten is an Assistant Professor at Boston University School of Social Work in Boston, Massachusetts, and a licensed Clinical Social Worker. She also serves as an Officer in the Army National Guard. Dr. Wooten holds a Ph.D. Social Work from the University of Maryland, School of Social Work in Baltimore, Maryland.

Michael G. Rumsey

Dr. Michael G. Rumsey is Chief of the Personnel Assessment Research Unit of the U.S. Army Research Institute (ARI) for the Behavioral and Social Sciences in Arlington, Virginia. He also serves on the Editorial Board of the Journal *Military Psychology*. Dr. Rumsey holds a Ph.D. in Social Psychology from Purdue University in Indiana.

Dr. Peter F. Ramsberger

Dr. Nikki R. Wooten

Dr. Michael G. Rumsey

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
June 2012	Final	October 2006 – June 2012 1994

4. TITLE AND SUBTITLE
A HISTORY OF THE RESEARCH INTO METHODS FOR SELECTING AND CLASSIFYING U.S. ARMY PERSONNEL1917-2010

5a. CONTRACT NUMBER
DASW01-03-D-0015

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
N/A

6. EDITOR
PETER F. RAMSBERGER, NIKKI R. WOOTEN, MICHAEL G. RUMSEY

5d. PROJECT NUMBER
N/A

5e. TASK NUMBER
N/A

5f. WORK UNIT NUMBER
N/A

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U. S. Army Research Institute
for the Behavioral & Social Sciences
6000 6TH Street (Bldg. 1464 / Mail Stop 5610)
Fort Belvoir, VA 22060-5610

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Institute
for the Behavioral & Social Sciences
6000 6TH Street (Bldg. 1464 / Mail Stop 5610)
Fort Belvoir, VA 22060-5610

10. SPONSOR/MONITOR'S ACRONYM(S)
ARI

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
N/A

12. DISTRIBUTION/AVAILABILITY STATEMENT: Distribution Statement A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Army scientists have been in the forefront of the behavioral and social science research that underlies the selection and classification of personnel for generations. In many ways, the history of the United States Army's selection and classification research is the history of selection and classification testing in general. The research in selection and classification described in this volume is a case study in the development of ideas, the conduct of basic and applied research, and its continued successful application to significant and critical personnel issues. The research is placed not only in a historical context, but also in the context of current day Army selection and classification procedures.This book documents military research contributions, over a 90 year time span, leading to the important decisions that the Army makes today about whom to select and what jobs to assign to them. It also tells the story of the scientific contributions Army researchers have made over the years to the personnel testing literature and practice.

15. SUBJECT TERMS
Army Personnel, Personnel Selection and Classification, Personnel - Research, Army Research, History-Army Personnel, Personnel-Testing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					Dorothy Young
a. REPORT	b. ABSTRACT	c. THIS PAGE	Unlimited Unclassified	260	19b. TELEPHONE NUMBER
Unclassified	Unclassified	Unclassified			703-545-2316